
JMIR AI

Volume 2 (2023) ISSN 2817-1705 Editors-in-Chief: Khaled El Emam, Bradley Malin, PhD

Contents

Editorial

- Reporting and Methodological Observations on Prognostic and Diagnostic Machine Learning Studies
([e47995](#))
Khaled El Emam, William Klement, Bradley Malin. 6

Tutorial

- Practical Considerations and Applied Examples of Cross-Validation for Model Development and Evaluation
in Health Care: Tutorial ([e49023](#))
Drew Wilimitis, Colin Walsh. 10

Reviews

- Strategies to Improve the Impact of Artificial Intelligence on Health Equity: Scoping Review ([e42936](#))
Carl Berdahl, Lawrence Baker, Sean Mann, Osonde Osoba, Federico Giroi. 26
- Application of a Comprehensive Evaluation Framework to COVID-19 Studies: Systematic Review of
Translational Aspects of Artificial Intelligence in Health Care ([e42313](#))
Aaron Casey, Saba Ansari, Bahareh Nakisa, Blair Kelly, Pieta Brown, Paul Cooper, Imran Muhammad, Steven Livingstone, Sandeep Reddy,
Ville-Petteri Makinen. 51
- Predicting Adherence to Behavior Change Support Systems Using Machine Learning: Systematic Review
([e46779](#))
Akon Ekpezu, Isaac Wiafe, Harri Oinas-Kukkonen. 64
- Machine Learning–Based Asthma Attack Prediction Models From Routinely Collected Electronic Health
Records: Systematic Scoping Review ([e46717](#))
Arif Budiarto, Kevin Tsang, Andrew Wilson, Aziz Sheikh, Syed Shah. 76
- The Application of Artificial Intelligence in Health Care Resource Allocation Before and During the COVID-19
Pandemic: Scoping Review ([e38397](#))
Hao Wu, Xiaoyu Lu, Hanyu Wang. 465

Original Papers

Forecasting Artificial Intelligence Trends in Health Care: Systematic International Patent Analysis (e47283) Stan Benjamins, Pranavsingh Dhunoo, Márton Görög, Bertalan Mesko.	41
Application of Artificial Intelligence to the Monitoring of Medication Adherence for Tuberculosis Treatment in Africa: Algorithm Development and Validation (e40167) Juliet Sekandi, Weili Shi, Ronghang Zhu, Patrick Kaggwa, Ernest Mwebaze, Sheng Li.	94
Effect of Benign Biopsy Findings on an Artificial Intelligence–Based Cancer Detector in Screening Mammography: Retrospective Case-Control Study (e48123) Athanasios Zouzos, Aleksandra Milovanovic, Karin Dembrower, Fredrik Strand.	105
Deep Learning Transformer Models for Building a Comprehensive and Real-time Trauma Observatory: Development and Validation Study (e40843) Gabrielle Chenais, Cédric Gil-Jardiné, Hélène Touchais, Marta Avalos Fernandez, Benjamin Contrand, Eric Tellier, Xavier Combes, Loick Bourdois, Philippe Revel, Emmanuel Lagarde.	124
Few-Shot Learning for Clinical Natural Language Processing Using Siamese Neural Networks: Algorithm Development and Validation Study (e44293) David Oniani, Premkumar Chandrasekar, Sonish Sivarajkumar, Yanshan Wang.	138
Extraction of Radiological Characteristics From Free-Text Imaging Reports Using Natural Language Processing Among Patients With Ischemic and Hemorrhagic Stroke: Algorithm Development and Validation (e42884) Enshuo Hsu, Abdulaziz Bako, Thomas Potter, Alan Pan, Gavin Britz, Jonika Tannous, Farhaan Vahidy.	147
Natural Language Processing for Clinical Laboratory Data Repository Systems: Implementation and Evaluation for Respiratory Viruses (e44835) Elham Dolatabadi, Branson Chen, Sarah Buchan, Alex Austin, Mahmoud Azimaee, Allison McGeer, Samira Mubareka, Jeffrey Kwong.	159
Developing an Inpatient Electronic Medical Record Phenotype for Hospital-Acquired Pressure Injuries: Case Study Using Natural Language Processing Models (e41264) Elvira Nurmambetova, Jie Pan, Zilong Zhang, Guosong Wu, Seungwon Lee, Danielle Southern, Elliot Martin, Chester Ho, Yuan Xu, Cathy Eastwood.	170
Automated Identification of Aspirin-Exacerbated Respiratory Disease Using Natural Language Processing and Machine Learning: Algorithm Development and Evaluation Study (e44191) Thanai Pongdee, Nicholas Larson, Rohit Divekar, Suzette Bielinski, Hongfang Liu, Sungrim Moon.	185
Extractive Clinical Question-Answering With Multianswer and Multifocus Questions: Data Set Development and Evaluation Study (e41818) Sungrim Moon, Huan He, Heling Jia, Hongfang Liu, Jungwei Fan.	192
Momentary Depressive Feeling Detection Using X (Formerly Twitter) Data: Contextual Language Approach (e49531) Ali Jamali, Corinne Berger, Raymond Spiteri.	202
Developing Ethics and Equity Principles, Terms, and Engagement Tools to Advance Health Equity and Researcher Diversity in AI and Machine Learning: Modified Delphi Approach (e52888) Rachele Hendricks-Sturup, Malaika Simmons, Shilo Anders, Kammarauche Aneni, Ellen Wright Clayton, Joseph Coco, Benjamin Collins, Elizabeth Heitman, Sajid Hussain, Karuna Joshi, Josh Lemieux, Laurie Lovett Novak, Daniel Rubin, Anil Shanker, Talitha Washington, Gabriella Waters, Joyce Webb Harris, Rui Yin, Teresa Wagner, Zhijun Yin, Bradley Malin.	214

A Scalable Radiomics- and Natural Language Processing–Based Machine Learning Pipeline to Distinguish Between Painful and Painless Thoracic Spinal Bone Metastases: Retrospective Algorithm Development and Validation Study (e44779)	
Hossein Naseri, Sonia Skamene, Marwan Tolba, Mame Faye, Paul Ramia, Julia Khriugian, Marc David, John Kildea.	226
Detecting Ground Glass Opacity Features in Patients With Lung Cancer: Automated Extraction and Longitudinal Analysis via Deep Learning–Based Natural Language Processing (e44537)	
Kyeryoung Lee, Zongzhi Liu, Urmila Chandran, Iftekhar Kalsekar, Balaji Laxmanan, Mitchell Higashi, Tomi Jun, Meng Ma, Minghao Li, Yun Mai, Christopher Gilman, Tongyu Wang, Lei Ai, Parag Aggarwal, Qi Pan, William Oh, Gustavo Stolovitzky, Eric Schadt, Xiaoyan Wang.	239
A Trainable Open-Source Machine Learning Accelerometer Activity Recognition Toolbox: Deep Learning Approach (e42337)	
Fluri Wieland, Claudio Nigg.	254
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 (e40973)	
Tharshini Jeyakumar, Sarah Younus, Melody Zhang, Megan Clare, Rebecca Charow, Inaara Karsan, Azra Dhalla, Dalia Al-Mouaswas, Jillian Scandiffio, Justin Aling, Mohammad Salhia, Nadim Lalani, Scott Overholt, David Wiljer.	265
Artificial Intelligence in Health Care—Understanding Patient Information Needs and Designing Comprehensible Transparency: Qualitative Study (e46487)	
Renee Robinson, Cara Liday, Sarah Lee, Ishan Williams, Melanie Wright, Sungjoon An, Elaine Nguyen.	282
Artificial Intelligence–Enabled Software Prototype to Inform Opioid Pharmacovigilance From Electronic Health Records: Development and Usability Study (e45000)	
Alfred Sorbello, Syed Haque, Rashedul Hasan, Richard Jermyn, Ahmad Hussein, Alex Vega, Krzysztof Zembrzuski, Anna Ripple, Mitra Ahadpour.	293
Self-Supervised Electroencephalogram Representation Learning for Automatic Sleep Staging: Model Development and Evaluation Study (e46769)	
Chaoqi Yang, Cao Xiao, M Westover, Jimeng Sun.	306
Enabling Early Health Care Intervention by Detecting Depression in Users of Web-Based Forums using Language Models: Longitudinal Analysis and Evaluation (e41205)	
David Owen, Dimosthenis Antypas, Athanasios Hassoulas, Antonio Pardiñas, Luis Espinosa-Anke, Jose Collados.	320
Machine Learning for the Prediction of Procedural Case Durations Developed Using a Large Multicenter Database: Algorithm Development and Validation Study (e44909)	
Samir Kendale, Andrew Bishara, Michael Burns, Stuart Solomon, Matthew Corriere, Michael Mathis.	345
Association of Health Care Work With Anxiety and Depression During the COVID-19 Pandemic: Structural Topic Modeling Study (e47223)	
Matteo Malgaroli, Emily Tseng, Thomas Hull, Emma Jennings, Tanzeem Choudhury, Naomi Simon.	361
Assessing Elevated Blood Glucose Levels Through Blood Glucose Evaluation and Monitoring Using Machine Learning and Wearable Photoplethysmography Sensors: Algorithm Development and Validation (e48340)	
Bohan Shi, Satvinder Dhaliwal, Marcus Soo, Cheri Chan, Jocelin Wong, Natalie Lam, Entong Zhou, Vivien Paitimusa, Kum Loke, Joel Chin, Mei Chua, Kathy Liaw, Amos Lim, Fadil Insyirah, Shih-Cheng Yen, Arthur Tay, Seng Ang.	375
Insights on the Current State and Future Outlook of AI in Health Care: Expert Interview Study (e47353)	
Pia Hummelsberger, Timo Koch, Sabrina Rauh, Julia Dorn, Eva Lermer, Martina Raue, Matthias Hudecek, Andreas Schicho, Errol Colak, Marzyeh Ghassemi, Susanne Gaube.	391
Machine Learning Models Versus the National Early Warning Score System for Predicting Deterioration: Retrospective Cohort Study in the United Arab Emirates (e45257)	
Hazem Lashen, Terrence St John, Y Almallah, Madhu Sasidhar, Farah Shamout.	407

Using Conversational AI to Facilitate Mental Health Assessments and Improve Clinical Efficiency Within Psychotherapy Services: Real-World Observational Study (e44358)	
Max Rollwage, Johanna Habicht, Keno Juechems, Ben Carrington, Sruthi Viswanathan, Mona Stylianou, Tobias Hauser, Ross Harper.	423
The Evolution of Artificial Intelligence in Biomedicine: Bibliometric Analysis (e45770)	
Jiasheng Gu, Chongyang Gao, Lili Wang.	436
Association Between Online Reviews of Substance Use Disorder Treatment Facilities and Drug-Induced Mortality Rates: Cross-Sectional Analysis (e46317)	
Matthew Abrams, Raina Merchant, Zachary Meisel, Arthur Pelullo, Sharath Chandra Guntuku, Anish Agarwal.	454
Predicting Treatment Interruption Among People Living With HIV in Nigeria: Machine Learning Approach (e44432)	
Matthew-David Ogbachie, Christa Fischer Walker, Mu-Tien Lee, Amina Abba Gana, Abimbola Oduola, Augustine Idemudia, Matthew Etor, Emily Harris, Jessica Stephens, Xiaoming Gao, Pai-Lien Chen, Navindra Persaud.	478
Machine Learning–Based Time in Patterns for Blood Glucose Fluctuation Pattern Recognition in Type 1 Diabetes Management: Development and Validation Study (e45450)	
Nicholas Chan, Weizi Li, Theingi Aung, Eghosa Bazuaye, Rosa Montero.	487
An Assessment of How Clinicians and Staff Members Use a Diabetes Artificial Intelligence Prediction Tool: Mixed Methods Study (e45032)	
Winston Liaw, Yessenia Ramos Silva, Erica Soltero, Alex Krist, Angela Stotts.	505
Identifying the Question Similarity of Regulatory Documents in the Pharmaceutical Industry by Using the Recognizing Question Entailment System: Evaluation Study (e43483)	
Nidhi Saraswat, Chuqin Li, Min Jiang.	518
Physicians' and Machine Learning Researchers' Perspectives on Ethical Issues in the Early Development of Clinical Machine Learning Tools: Qualitative Interview Study (e47449)	
Jane Kim, Katie Ryan, Max Kasun, Justin Hogg, Laura Dunn, Laura Roberts.	532
Real-Time Classification of Causes of Death Using AI: Sensitivity Analysis (e40965)	
Patrícia Pita Ferreira, Diogo Godinho Simões, Constança Pinto de Carvalho, Francisco Duarte, Eugénia Fernandes, Pedro Casaca Carvalho, José Loff, Ana Soares, Maria Albuquerque, Pedro Pinto-Leite, André Peralta-Santos.	545
Determinants of Intravenous Infusion Longevity and Infusion Failure via a Nonlinear Model Analysis of Smart Pump Event Logs: Retrospective Study (e48628)	
Arash Kia, James Waterson, Norma Bargary, Stuart Rolt, Kevin Burke, Jeremy Robertson, Samuel Garcia, Alessio Benavoli, David Bergström.	5 6 4
Predicting Patient Mortality for Earlier Palliative Care Identification in Medicare Advantage Plans: Features of a Machine Learning Model (e42253)	
Anne Bowers, Chelsea Drake, Alexi Makarkin, Robert Monzyk, Biswajit Maity, Andrew Telle.	575
Deep Learning to Detect Pancreatic Cystic Lesions on Abdominal Computed Tomography Scans: Development and Validation Study (e40702)	
Maria Duh, Neus Torra-Ferrer, Meritxell Riera-Marín, Dídac Cumelles, Júlia Rodríguez-Comas, Javier García López, M ^a Fernández Planas.	5 8 8
Patient Embeddings From Diagnosis Codes for Health Care Prediction Tasks: Pat2Vec Machine Learning Framework (e40755)	
Edgar Steiger, Lars Kroll.	597



Prediction of Chronic Stress and Protective Factors in Adults: Development of an Interpretable Prediction Model Based on XGBoost and SHAP Using National Cross-sectional DEGS1 Data (e41868)	
Arezoo Bozorgmehr, Birgitta Weltermann.	613

Viewpoint

Artificial Intelligence Algorithms in Health Care: Is the Current Food and Drug Administration Regulation Sufficient? (e42940)	
Meghavi Mashar, Shreya Chawla, Fangyue Chen, Baker Lubwama, Kyle Patel, Mihir Kelshiker, Patrik Bachtiger, Nicholas Peters.	114

Editorial

Reporting and Methodological Observations on Prognostic and Diagnostic Machine Learning Studies

Khaled El Emam^{1,2}, BEng, PhD; William Klement^{1,2}, PhD; Bradley Malin³, MSc, PhD

¹School of Epidemiology and Public Health, University of Ottawa, Ottawa, ON, Canada

²CHEO Research Institute, Ottawa, ON, Canada

³Department of Biomedical Informatics, Vanderbilt University, Nashville, TN, United States

Corresponding Author:

Khaled El Emam, BEng, PhD

School of Epidemiology and Public Health

University of Ottawa

401 Smyth Rd

Ottawa, ON, K1H 8L1

Canada

Phone: 1 6137975412

Email: kelemam@ehealthinformation.ca

Abstract

Common reporting and methodological patterns were observed from the peer reviews of prognostic and diagnostic machine learning modeling studies submitted to JMIR AI. In this editorial, we summarized some key observations to inform future studies and their reporting.

(JMIR AI 2023;2:e47995) doi:[10.2196/47995](https://doi.org/10.2196/47995)

KEYWORDS

reporting guidelines; machine learning; modeling studies; prognostic studies; methodological observations; diagnostic studies; ML models

Introduction

The JMIR AI journal was launched at the beginning of 2022. During that first year, many of the papers submitted to the journal reported on prognostic studies that applied machine learning (ML) models. In this editorial update, we wish to highlight common patterns that were observed from the comments of the peer reviewers. Our objective in publishing this editorial is to inform authors about specific issues that should be documented and provide information about common methodological problems that can be avoided. Since these observations can help improve articles submitted to the journal, authors will benefit both in terms of acceptance rates and turnaround times for publication decisions. Furthermore, these observations may be of value to the broader ML community to inform the reporting of their studies. They are not intended to be comprehensive reporting guidelines but focus specifically on our observations with journal submissions.

We examined reviewers' comments for papers submitted to JMIR AI over the entirety of 2022 (irrespective of their eventual publication decision). This included all papers remaining under review. We focused solely on papers that presented prognostic and diagnostic models using ML modeling techniques. The

most common suggestions or critiques raised by reviewers were identified by counting observations in the reviewer comments. It was recognized that, at times, reviewers' comments covered multiple overlapping issues or implied an issue without stating it completely. As a consequence, some judgment by us was required to decide which reviewer observations should be included in this update.

Reporting and Methodological Observations

The Degrees of Limitations

In some instances, there was a methodological weakness in the study. If this is raised by a reviewer, there is a tendency for authors to mention this issue in the "Limitations" section of the manuscript, rather than address it in the study itself. However, some weaknesses are not just standard limitations but affect the meaningfulness of the modeling that was performed and whether valid conclusions can be drawn from it. Not all weaknesses will be considered acceptable limitations, some of which we highlight throughout this article.

The limitations communicated in a manuscript present shortcomings due to practical or theoretical constraints presented to the model or algorithm, in which case it is anticipated that the constraints are out of the control of the authors and may inspire future research directions. As a hypothetical example, imagine that tissue samples are collected from donor lungs prior to lung transplantation, and a researcher subsequently develops a prognostic molecular test to predict if an adverse event will occur within the first 72 hours after lung transplantation surgery. This test is fundamentally limited to the molecular makeup of the donor because it neglects to consider the immunological response of the transplant recipient toward the prediction. In practice, surgical constraints prevent the collection of tissue samples immediately after transplantation.

In contrast, limitations by choice reflect decisions made in the scope, focus, methodology, and possibly aims of the study that can result in weaknesses that may be deemed unnecessary. The latter type of limitations needs to be addressed in the (re)submitted manuscript, which may require further analysis and rework. Of course, some judgment is necessary to distinguish between the two types of limitations, but the default of adding critical weaknesses to the “Limitations” section of a study report is not recommended.

Documenting Reasons and Impacts of Data Sampling

Some studies start with a very large number of observations but end up using only a small proportion in the study. In many cases, the reduction in sample size is not an artifact of a random process. In such a case, it is possible that the authors have induced a selection bias in the data [1,2]. For example, if there are 1000 patient records with a particular diagnosis in a health care organization that meet the inclusion criteria, but only 500 are used in the study, how, if at all, does the subset differ from the initial larger group?

In some cases, missingness is a reason why many patients are excluded from an analysis. It is plausible that missing values of certain variables, which may include the outcome itself, may be correlated with specific groups of patients. Thus, the authors should try to explain how missingness affects patient characteristics. Could the patients with missing values be less severe cases and therefore the data set used to train a prognostic model consists of healthier patients? And, if this is the case, is the trained model capable of generalizing to the broader population when it is applied in practice?

Avoiding Data Leakage

It is important to be cognizant of data leakage in model evaluation; otherwise, optimistic results may be obtained. An example of leakage is when there are multiple observations per patient distributed across the training and testing subsets of the data set. Effectively, information about the same patient may be included in both the training and testing data sets. Because the observations in the training and testing data sets are likely to be correlated, the error rate may be optimistic. Special care must be exercised to ensure that such leakage does not compromise the results of the analysis [2].

From a reporting perspective, authors should clarify if there are repeated or correlated observations, as well as the actions taken to avoid data leakage [3].

Reporting Missingness and How It Is Handled

It is important to indicate how many observations were missing for each variable included in model building. If specific actions were performed to handle missingness, then these should be stated as well. For example, authors should report if a complete case analysis or a specific type of imputation was performed [3-5]. Moreover, if imputation methods are applied, then the affected variables and the imputation methods need to be reported and their parameterizations need to be described [4,6].

Justifying the Choice of ML Model(s)

Justification of ML modeling techniques is a somewhat common reviewer comment regarding deficiencies in a manuscript. Some studies compare the performance of different types of ML models. In such situations, the selection of ML models should be justified [7-9].

Using logistic regression as a baseline is often a reasonable choice as it is a commonly used modeling method [10]. A recent systematic review showed that logistic regression performance is comparable to the use of ML models for clinical prediction workloads [11]. Therefore, it represents a realistic baseline workload. The choice of other methods should be justified. For example, it may be the case that an ML model is selected because it is commonly relied upon by the academic community or is a standard in practice. Moreover, it may be the case that a particular method is considered state of the art.

Reporting Hyperparameter Tuning Methodology and Results

An ML algorithm is typically controlled by a collection of hyperparameters that influence how learning takes place. Authors should describe if any hyperparameter tuning was performed or if and what default parameters were used. If hyperparameter tuning was performed, then an explanation should indicate which method was applied (eg, grid search or Bayesian optimization), as well as what loss function was relied upon. If one or more models are being reported upon, then the final parameters should be included in the supplementary materials. An exception would be reasonable in the context of a simulation where thousands of models may be trained. In this case, a method indicating how the models are generated should be detailed to ensure reproducibility [3,7].

The method for evaluating the performance of the tuned model should also be described. For example, nested cross-validation would allow the performance to be computed on the tuned models. Then, the final set of hyperparameters is determined from a follow-up k-fold cross-validation, and these latter ones should be reported [8,9].

Documenting the Decision Threshold

Studies that use classification or regression, where a decision threshold maps the classification scores to a class or category, are common. The decision threshold can have a big impact on the performance of the model [12,13], and the relative cost of incorrect decisions. The often-used default threshold of 0.5 is

not always a good choice. Documentation of the threshold and justification for the value selected are necessary to enable the reader to properly interpret the model performance.

Conclusions

While this summary pertains to prognostic and diagnostic models mostly for structured data, many of the points are

relevant for other types of data modalities (eg, image processing). Moreover, it should be recognized that the observations covered in this editorial are not exhaustive as there are other subtle issues that are highlighted by reviewers for specific studies. Nonetheless, adhering to the reporting recommendations and methodological considerations indicated above will be beneficial for *JMIR AI* submissions.

Conflicts of Interest

KE and BM are Editors-in-Chief of *JMIR AI* at the time of this publication.

References

1. Hoens TR, Chawla NV. Chapter 3 Imbalanced datasets: from sampling to classifiers. In: He H, Ma Y, editors. *Imbalanced Learning: Foundations, Algorithms, and Applications*. Hoboken, NJ: John Wiley & Sons, Inc; 2013:43-59.
2. Kaufman S, Rosset S, Perlich C, Stitelman O. Leakage in data mining. *ACM Trans Knowl Discov Data* 2012 Dec 18;6(4):1-21. [doi: [10.1145/2382577.2382579](https://doi.org/10.1145/2382577.2382579)]
3. Cabitza F, Campagner A. The need to separate the wheat from the chaff in medical informatics: introducing a comprehensive checklist for the (self)-assessment of medical AI studies. *Int J Med Inform* 2021 Sep;153:104510 [FREE Full text] [doi: [10.1016/j.ijmedinf.2021.104510](https://doi.org/10.1016/j.ijmedinf.2021.104510)] [Medline: [34108105](https://pubmed.ncbi.nlm.nih.gov/34108105/)]
4. Jadhav A, Pramod D, Ramanathan K. Comparison of performance of data imputation methods for numeric dataset. *Appl Artif Intell* 2019 Jul 04;33(10):913-933. [doi: [10.1080/08839514.2019.1637138](https://doi.org/10.1080/08839514.2019.1637138)]
5. Zhao Y, Long Q. Variable selection in the presence of missing data: imputation-based methods. *Wiley Interdiscip Rev Comput Stat* 2017 May 24;9(5):e1402 [FREE Full text] [doi: [10.1002/wics.1402](https://doi.org/10.1002/wics.1402)] [Medline: [29085552](https://pubmed.ncbi.nlm.nih.gov/29085552/)]
6. Lu JH, Callahan A, Patel BS, Morse KE, Dash D, Shah NH. Low adherence to existing model reporting guidelines by commonly used clinical prediction models. medRxiv. Preprint posted online July 23, 2021. [doi: [10.1101/2021.07.21.21260282](https://doi.org/10.1101/2021.07.21.21260282)]
7. Scott I, Carter S, Coiera E. Clinician checklist for assessing suitability of machine learning applications in healthcare. *BMJ Health Care Inform* 2021 Feb 05;28(1):e100251 [FREE Full text] [doi: [10.1136/bmjhci-2020-100251](https://doi.org/10.1136/bmjhci-2020-100251)] [Medline: [33547086](https://pubmed.ncbi.nlm.nih.gov/33547086/)]
8. van Smeden M, Heinze G, Van Calster B, Asselbergs FW, Vardas PE, Bruining N, et al. Critical appraisal of artificial intelligence-based prediction models for cardiovascular disease. *Eur Heart J* 2022 Aug 14;43(31):2921-2930 [FREE Full text] [doi: [10.1093/eurheartj/ehac238](https://doi.org/10.1093/eurheartj/ehac238)] [Medline: [35639667](https://pubmed.ncbi.nlm.nih.gov/35639667/)]
9. Luo W, Phung D, Tran T, Gupta S, Rana S, Karmakar C, et al. Guidelines for developing and reporting machine learning predictive models in biomedical research: a multidisciplinary view. *J Med Internet Res* 2016 Dec 16;18(12):e323 [FREE Full text] [doi: [10.2196/jmir.5870](https://doi.org/10.2196/jmir.5870)] [Medline: [27986644](https://pubmed.ncbi.nlm.nih.gov/27986644/)]
10. Collins GS, Reitsma JB, Altman DG, Moons KGM. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): the TRIPOD statement. *BMJ* 2015 Jan 07;350:g7594 [FREE Full text] [doi: [10.1136/bmj.g7594](https://doi.org/10.1136/bmj.g7594)] [Medline: [25569120](https://pubmed.ncbi.nlm.nih.gov/25569120/)]
11. Christodoulou E, Ma J, Collins GS, Steyerberg EW, Verbakel JY, Van Calster B. A systematic review shows no performance benefit of machine learning over logistic regression for clinical prediction models. *J Clin Epidemiol* 2019 Jun;110:12-22. [doi: [10.1016/j.jclinepi.2019.02.004](https://doi.org/10.1016/j.jclinepi.2019.02.004)] [Medline: [30763612](https://pubmed.ncbi.nlm.nih.gov/30763612/)]
12. Sengupta PP, Shrestha S, Berthon B, Messas E, Donal E, Tison GH, et al. Proposed Requirements for Cardiovascular Imaging-Related Machine Learning Evaluation (PRIME): a checklist: reviewed by the American College of Cardiology Healthcare Innovation Council. *JACC Cardiovasc Imaging* 2020 Sep;13(9):2017-2035 [FREE Full text] [doi: [10.1016/j.jcmg.2020.07.015](https://doi.org/10.1016/j.jcmg.2020.07.015)] [Medline: [32912474](https://pubmed.ncbi.nlm.nih.gov/32912474/)]
13. Hernandez-Boussard T, Bozkurt S, Ioannidis JPA, Shah NH. MINIMAR (MINimum Information for Medical AI Reporting): developing reporting standards for artificial intelligence in health care. *J Am Med Inform Assoc* 2020 Dec 09;27(12):2011-2015 [FREE Full text] [doi: [10.1093/jamia/ocaa088](https://doi.org/10.1093/jamia/ocaa088)] [Medline: [32594179](https://pubmed.ncbi.nlm.nih.gov/32594179/)]

Abbreviations

ML: machine learning

Edited by T Leung; submitted 07.04.23; this is a non-peer-reviewed article; accepted 07.04.23; published 28.04.23.

Please cite as:

El Emam K, Klement W, Malin B

Reporting and Methodological Observations on Prognostic and Diagnostic Machine Learning Studies

JMIR AI 2023;2:e47995

URL: <https://ai.jmir.org/2023/1/e47995>

doi: [10.2196/47995](https://doi.org/10.2196/47995)

PMID: [32148429](https://pubmed.ncbi.nlm.nih.gov/32148429/)

©Khaled El Emam, William Klement, Bradley Malin. Originally published in JMIR AI (<https://ai.jmir.org>), 28.04.2023. This is an open-access article distributed under the terms of the Creative Commons Attribution License (<https://creativecommons.org/licenses/by/4.0/>), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR AI, is properly cited. The complete bibliographic information, a link to the original publication on <https://www.ai.jmir.org/>, as well as this copyright and license information must be included.

Tutorial

Practical Considerations and Applied Examples of Cross-Validation for Model Development and Evaluation in Health Care: Tutorial

Drew Wilimitis¹, BS; Colin G Walsh¹, MA, MD

Vanderbilt University Medical Center, Vanderbilt University, Nashville, TN, United States

Corresponding Author:

Colin G Walsh, MA, MD

Vanderbilt University Medical Center

Vanderbilt University

2525 W End Ave, Suite 1475

Nashville, TN, 37203

United States

Phone: 1 6159365684

Email: colin.walsh@vumc.org

Abstract

Cross-validation remains a popular means of developing and validating artificial intelligence for health care. Numerous subtypes of cross-validation exist. Although tutorials on this validation strategy have been published and some with applied examples, we present here a practical tutorial comparing multiple forms of cross-validation using a widely accessible, real-world electronic health care data set: Medical Information Mart for Intensive Care-III (MIMIC-III). This tutorial explored methods such as K-fold cross-validation and nested cross-validation, highlighting their advantages and disadvantages across 2 common predictive modeling use cases: classification (mortality) and regression (length of stay). We aimed to provide readers with reproducible notebooks and best practices for modeling with electronic health care data. We also described sets of useful recommendations as we demonstrated that nested cross-validation reduces optimistic bias but comes with additional computational challenges. This tutorial might improve the community's understanding of these important methods while catalyzing the modeling community to apply these guides directly in their work using the published code.

(*JMIR AI* 2023;2:e49023) doi:[10.2196/49023](https://doi.org/10.2196/49023)

KEYWORDS

predictive modeling; cross-validation; tutorial; model development; risk detection; clinical decision-making; electronic health care; eHealth data; health care data; data validation; artificial intelligence; AI

Background

By learning complex statistical relationships from historical data, predictive models enable automated and scalable risk detection and prognostication, which might inform clinical decision-making. Although relatively few have been implemented in clinical use compared with the number developed, predictive models are increasingly being deployed and tested in clinical trials. The importance of predictive modeling is on the rise, with increasing attention from regulatory bodies such as the US Food and Drug Administration. Efforts to standardize the steps in model development and validation include statements such as transparent reporting of a multivariable prediction model for individual prognosis or diagnosis and multiple published guidelines on deployment and governance [1-3]. However, the mode in a critical step in model development, the validation strategy, remains a simple “holdout”

or “test-train split,” which has been shown to introduce bias, fail to generalize, and hinder clinical utility [4-6].



Broadly, validation consists of either internal validation, which should be reported alongside model development, or external validation, in which a developed model is tested in an unseen data set in a new setting [7,8]. A newer concept of “internal-external” validation has also been suggested for studies with multisite data [9]. Most published models evaluate performance metrics by splitting the available data set into an independent “holdout” or “test set,” consisting of unforeseen samples excluded from model training. Such held-out sets are often selected randomly, for example, “80% training and 20% testing,” from the data in the original model development setting. In contrast to holdout validation, cross-validation and resampling methods such as bootstrapping can be used to produce less biased estimates of the true out-of-sample performance (ie, the ability to generalize to new samples). Although cross-validation is a widely used and extensively studied statistical method,

many variations of cross-validation exist with respective strengths and weaknesses, distinct use cases for model development and performance estimation that are often misapplied, and domain-specific considerations necessary for effective health care implementation [10,11].

Cross-validation surveys with practical examples, such as those involving microarray and neurologic data, have been published [12,13]. However, gaps in comprehensive tutorials including complete codesets with relevant tutorial data are less well disseminated. Tutorials that move beyond simulated data or laboratory-based samples to real-world health care data sets might add to the understanding of these important methods while catalyzing the modeling community to apply these guides directly in their work using the published code.

The intent of this tutorial is to define and compare means of cross-validation using representative, accessible data based in the well-known and well-studied Medical Information Mart for Intensive Care-III (MIMIC-III) data set [14]. All cross-validation modeling experiments and preprocessing codes will be provided through reproducible notebooks that will further guide readers through the comparisons and concepts introduced [15]. Best practices and common missteps, particularly in modeling with electronic health care data, will be emphasized.

Overview and Major Types of Cross-Validation

The goal of supervised learning is to use a data set with known labels, $D(X_i, Y_i)$, to produce a model  that accurately predicts the true labels of unforeseen “test” samples Y_i .  must learn robust relationships between the covariate features $[X_1 \dots X_n]$ and the outcome of interest. The model is considered a statistical estimator of Y_i because the model prediction is calculated from the available training data, and Y_i is a random variable with an unknown probability distribution. Given finite data samples with inherent statistical noise, the generalization or “test” error of this estimator will be imperfect. Assuming the true label to be a continuous outcome, we can decompose the mean-squared error of the learned model into 2 fundamental sources of error: *bias* and *variance*, formalized in the equation below by the first and second terms on the right hand side, respectively. The s^2 term represents irreducible, independent, and identically distributed error terms attributed to noise in the training data set.



(1)

Understanding the tradeoff between *bias* and *variance* is necessary to develop useful predictive models in health care. Bias can be thought of as the model’s inability to discern complex statistical patterns associated with true test labels. Variance is the additional error owing to the model mistakenly interpreting random fluctuations in the training data set as a robust predictive signal. Bias can often be reduced by increasing the complexity of the model (ie, if the model is *underfit*) in

hopes of uncovering deeper statistical patterns within the training data. However, the tradeoff between bias and variance then occurs, as more complex models are liable to *overfit* to random noise in the training data (thus increasing the variance). Model validation strategies such as cross-validation also have implications for the *bias-variance tradeoff*. Cross-validation generally relates to this tradeoff, as larger numbers of folds (smaller numbers of records per fold) tend toward higher variance and lower bias, whereas smaller numbers of folds tend toward higher bias and lower variance.

Before delving into the technical details and comparative advantages of specific cross-validation methods, it is imperative to emphasize that cross-validation was developed as a method to estimate the *expected* out-of-sample prediction error of a model learned from a set of training data. Machine learning developers have typically lacked access to external data sets (the gold standard that allows direct estimation of the true out-of-sample prediction error), and cross-validation offers an improvement over existing internal validation methods such as holdout validation. In contrast to parametric, model-specific methods such as Bayesian Information Criteria that rely on strict assumptions, cross-validation is nonparametric, compatible with any supervised learning algorithm, and directly estimates the primary measure of model validity—whether predictive performance generalizes to new data points. Cross-validation has become increasingly prominent for internal validation in health care given its flexibility with diverse and sophisticated learning algorithms and the advantage of using all available data for model evaluation and selection (compared with using a single holdout validation set). The use of cross-validation over holdout validation is particularly advantageous with health care data sets that are often comparatively small to moderately sized, costly to obtain, or restricted by privacy and regulatory concerns.

Cross-validation originated in the 1930s with K-fold cross-validation, the most common form of cross-validation, described in the 1960s [16,17]. In this form of cross-validation, the development data set is split into some number, k —often 5 or 10—parts, or “folds,” as described below ([Multimedia Appendix 1](#)). Several variations of this approach have since been described, each with its own advantages and disadvantages for clinical modeling.

Considerations for Clinical Prediction and Implications for Cross-Validation

Most published predictive modeling studies focus on the classification of binary outcomes; however, fewer models have been developed to predict continuous or ordinal variables. Clinical data, especially those in secondary use, for example, electronic health records (EHRs), are also typified by (1) irregular time-sampling, (2) inconsistent repeated measures, and (3) sparsity and rarity. Missingness, noise, and anomalous outlier values are additional complicating factors associated with EHR data. Although not uniquely relevant to cross-validation, appropriate ways of handling these data issues within model development and evaluation pipelines are covered in the applied demonstrations and available code accompanying this tutorial.

As health care delivery varies naturally and widely between individuals, real-world health care data such as EHRs usually contain irregularly and inconsistently sampled measures within and across individuals. This factor has significant implications for cross-validation and is described by the differences between *subject-wise* and *record-wise* cross-validations. Subject-wise cross-validation maintains identity across splits, such that an individual's set of events cannot exist in both training and testing simultaneously. In the record-wise cross-validation approach, data are split by event and not by individual. Record-wise cross-validation thereby increases the risk that the same individual will have events split across training and testing. A model might then achieve a spuriously high apparent performance simply by reidentifying the individual in testing based on highly similar inputs used in training.

Although the technical details and practical guidelines of subject-wise versus record-wise cross-validation are debated, the best approach depends on the specific use case, the number of records, the size of the data set, and the degree of correlation within subject records [18]. Developers should also consider the unit of modeling, that is, making a prediction for a given person versus a given health care encounter or event such as a prescription. In the former example, record-wise cross-validation might be adopted for diagnosis at a given clinical encounter, and subject-wise cross-validation would be favorable for prognosis over time [19]. In the latter examples, the training data might include multiple events per person, and the number of events will also vary across individuals. Cross-validation poses an additional potential benefit in that training and testing strategies might split individuals in such a way that models can be trained on some folds and then applied as inputs for ensembling in different folds without increasing the risk of overfitting or data leakage [20].

Many clinical outcomes subject to predictive modeling studies are rare at the health-system scale (eg, $\leq 1\%$ incidence). Although rare outcomes create modeling challenges out of scope for this tutorial, they also impact cross-validation. Randomly partitioning data sets into training and test splits often produces folds with various outcome rates and even folds with no outcome instances. For binary classification problems, stratified cross-validation ensures that outcome rates are equal across folds, and it is recommended for classification problems (and should be considered necessary for highly imbalanced classes) [20].

Major Steps in Using Cross-Validation

Dividing steps into development steps and validation steps eases interpretation. Development steps include data cleaning and preprocessing—a time-consuming but critical task given noisy and often invalid health care data from real-world sources—feature selection, classifier selection, hyperparameter tuning, and model refitting (Textbox 1). For brevity and scope,

classifier selection will not be covered in detail but is emphasized as an important step in any predictive modeling pipeline. As classifiers, broadly parametric or nonparametric, require different assumptions and themselves pose disparate advantages and disadvantages, it has become standard to test multiple classifiers in the same modeling study. Moreover, ensembles of these classifiers are increasingly developed given the rise in complexity, depth, and breadth of real-world health care data sets in the literature.

Model performance metrics inform validation steps and, when properly contextualized in the clinical use case, suggest key metrics on which to either optimize or evaluate model performance. A detailed discussion of model performance metrics remains outside the scope of this tutorial and has been covered in depth elsewhere. In brief, prediction models must be evaluated in terms of discrimination (ie, the ability to predict higher probabilities for individuals with the outcome) and calibration (a measure of similarity between predicted probabilities and the observed risk) [21]. Two methods to evaluate discrimination via the area under the receiver operator characteristic curve (AUROC) and the area under the precision-recall curve (AUPR) from cross-validation include (1) pooling: averaging test-fold results at each point on the receiver operating characteristic or precision-recall curve and (2) averaging: reporting the average AUROC and AUPR over each test-fold metric. Calibration can also be assessed analogously using metrics such as the Brier score, calibration slope, and intercept. We highlight methods that can be used for calibrating predictions along with cross-validation in a provided Jupyter notebook. Clinical usefulness based in decision analysis is the third major area of evaluation [22]. Usefulness bridges model performance to utility, for example, showing how a model might reduce cost or increase the measurement of quality of life.

In addition to cross-validation, bootstrapping is another resampling-based method used to provide more accurate estimates of the model generalization performance than holdout validation. Bootstrapping involves randomly sampling with replacement from the entire data set to generate a training set that will not include all original samples. A model is then fitted on the bootstrap training set and evaluated on a test set comprising the remaining unselected observations. This process is repeated several times, where the number of iterations is typically referred to as the number of bootstraps, and a CI is generated from the collection of out-of-sample (sometimes called “out of bag”) performance metrics. Traditional bootstrapping is referred to as out-of-bag bootstrapping, whereas further improvements include the “0.632” and “0.632+” methods that apply additional forms of bias adjustment [23,24]. More advanced resampling methods include bootstrap-based cross-validation, Monte Carlo holdout validation (or repeated holdout sampling), and repeated nested cross-validation [25,26].

Textbox 1. List of steps required for cross-validation.

1. *Data cleaning* (outside the loop)—Basic manipulation and feature engineering (converting data types, one-hot label encoding, etc) can and should be completed on the entire data set before beginning cross-validation.
2. *Feature scaling and imputation* (within the loop)—Imputation and feature scaling based on the other values in the data set (such as standardization via mean-centering or normalization) need to be completed only on the training set—which we call “within the loop” as a reference to use with nested cross-validation. This is necessary to reduce data leakage that can be caused if values in the test set are used to impute or scale the values in the training set.
3. *Feature selection* (within the loop)—To reduce overfitting and the detection of spurious correlations between the outcome and independent variables, feature selection should be completed only on the training fold (“inner loop” of nested cross-validation) and then applied and evaluated on the test fold.
4. *Model selection* (within the loop)—To mitigate optimistic bias, the comparison of different modeling algorithms (eg, random forest vs logistic regression) should also be completed separately from model evaluation.
5. *Model selection* (within the loop)—Optimizing hyperparameters can also be done simultaneously with classifier selection and should be used when identifying the best modeling algorithm.
6. *Evaluation*—Evaluation by way of “averaging” or “pooling” should be completed separately from using cross-validation for model selection to reduce optimism resulting from overfitting parameters to the data set used for evaluation.
7. *Model refitting*—To produce a final model trained with all the available data, one should learn the ideal feature selection and model selection parameters from cross-validation and then train the selected algorithm with these parameters using the entire data set. The model can then be ported outside the development setting to use for external validation or within production-grade systems.

Case Study in Cross-Validation: In-Hospital Mortality and Length of Stay Prediction

MIMIC-III represents a well-known deidentified data set based in intensive care at Beth Israel Deaconess for approximately 40,000 patients who received care from 2001 to 2012 [14]. It has been studied extensively, given its relative accessibility compared with most health care modeling studies using EHRs, in which privacy or challenges in at-scale deidentification prohibit data sharing with publication [27].

To demonstrate the key concepts in cross-validation, we selected 2 exemplary problems that typify predictive modeling studies: classification and regression. For this case study, in-hospital mortality prediction will represent the former, whereas length of stay prediction will represent the latter. The models will be developed and validated using multiple forms of cross-validation, including K-fold, stratified, repeated, repeated stratified, and nested cross-validation. We also applied bootstrap methods to generate CIs for estimated model performance.

From patient visit records, we derived time-invariant features such as age, sex, and race, along with binary features indicating the presence of prior diagnoses using 25 higher-order categories of International Classification of Diseases codes grouped into Clinical Classifications Software codes. In-hospital mortality was defined as a binary classification problem, where 1 indicated that mortality occurred at any point during the hospital visit and 0 otherwise. Length of stay was defined in days and used as a continuous outcome for a separate regression prediction problem.

Preprocessing included imputation of continuous features using the median, setting outlier age values to a maximum of 110 years, and standardization of all numerical features. We also applied a feature selection routine to select only the top 10 features available for in-hospital mortality prediction and either

the top 30 or 50 features (where the number of features was included as a hyperparameter) for length of stay prediction. Finally, in-hospital mortality prediction was classified using logistic regression and grid search over hyperparameters including least absolute shrinkage and selection operator (L1), Ridge (L2), and no penalization and a range of regularization values. Length of stay prediction was performed using random forest regression and grid search over hyperparameters including the number of estimators and maximum tree depth.

In accordance with the best practices outlined for cross-validation and model selection in Figure 1, we implemented a nested cross-validation approach that performed all hyperparameter tuning and model selection steps within the “inner” cross-validation loop. Theoretically, this should mitigate the source of optimistic bias introduced when cross-validation is used to tune model parameters on the same data used for model performance evaluation (ie, observed performance can be spuriously high owing to randomness in the data and the learning algorithm) [28,29]. This source of bias in the estimated model performance can be considered as a type of overfitting in the model selection procedure [30].

To empirically evaluate whether nested cross-validation produces more accurate performance estimates than nonnested cross-validation, we compared the nested cross-validation with nonnested cross-validation used simultaneously for model selection and model evaluation. For nonnested cross-validation methods, we evaluated the performance of each set of model tuning configurations (eg, models trained with varied hyperparameters) on the test fold at each cross-validation split. After repeating this procedure for each split within the given cross-validation method, we computed the average performance over all test folds for each model tuning configuration. The model parameters with the best average performance over the cross-validation test folds were then selected. The performance of this optimal set of hyperparameters was then reported as an estimate of true out-of-sample performance.

To demonstrate the optimism that can result from improperly applied model validation strategies (ie, simultaneously applying nonnested cross-validation methods for both model selection and model evaluation), we evaluated the accuracy of the estimated true test performance when using various cross-validation methods. We performed this by randomly splitting the data set into an 80% (32,897/41,121) sample used for cross-validation and a 20% (8224/41,121) withheld validation sample. We used a holdout setup to simulate ground truth in the absence of a naturally bounded holdout (eg, by site or clinical setting) in the MIMIC data. We then compared the

best model performance reported from cross-validation with the performance of that model when predicting on the held-out validation set (Figure 2).

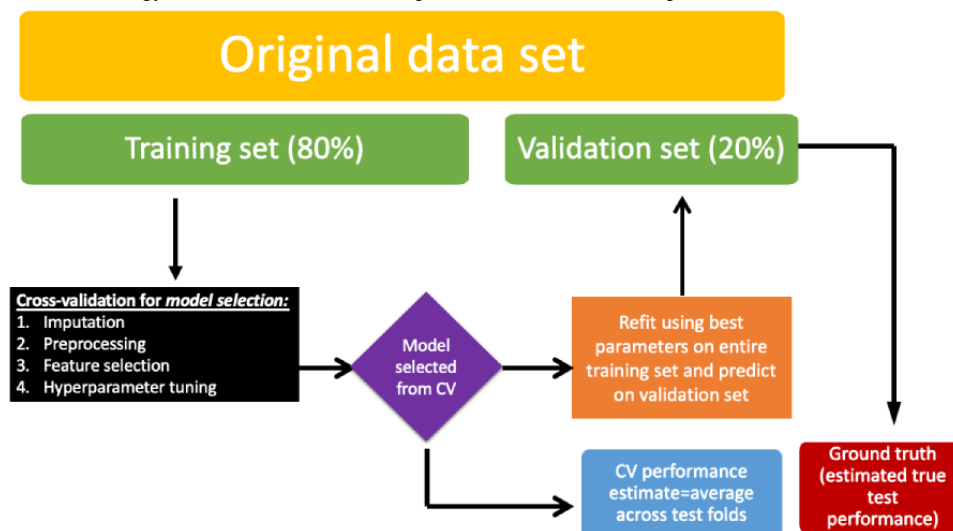
Performance measures will include discrimination metrics such as the AUROC and AUPR. For length of stay regression modeling, we adopted mean absolute error and median absolute error as the primary performance metrics. Computational time, a pragmatic concern affecting many modeling experiments, will also be compared across cross-validation methods for both prediction outcomes.

Figure 1. Pseudocode for nested cross-validation algorithm with model tuning.

```

Algorithm 1: Nested cross-validation with model tuning
Input: k1 (number of outer folds), k2 (number of inner folds), X (data set with features and outcome labels), M (prediction model algorithm), Params (set of model parameter specifications)
Apply K-fold CV to split X into k1 folds
for i = 1 to k1: # outer loop
    Let Xi_test be the ith fold # contains 1/k1 of the samples in X
    Let the remaining samples be Xi_train
    X_train_outer = Xi_train
    X_test_outer = Xi_test
    Apply K-fold CV to split X_train_outer into k2 folds
    for j = 1 to k2: # inner loop
        Let Xj_test be the jth fold # contains 1/k2 of the samples in X_train_outer
        Let the remaining samples be Xj_train
        X_train_inner = Xj_train
        X_test_inner = Xj_test
        for param_set in Params: # apply all tuning steps within inner loop
            Fit model M to X_train_inner with parameters in param_set
            Compute performance of M on X_test_inner
        end
    Let P* be the parameters with the best mean score over inner loop test sets
    Fit model M to X_train_outer with parameter set P*
    Compute performance of M on X_test_outer # evaluation performed on outer loop
end
    
```

Figure 2. Diagram of the methodology for cross-validation (CV) optimism error estimation experiment.



Case Study Results

Cohort Description

After applying the exclusion criteria, our cohort included 41,121 hospital visits that comprised 71.63% (29,457/41,121) of White patients and 55.92% (22,996/41,121) of male patients. Mortality was observed for 4320 patient visits (4320/41,121, 10.51% of the total cohort). The length of stay (days) did not vary across different demographic groups, whereas the mean age of patients with in-hospital mortality (68.7, SD 15.0 y) was greater than those without mortality events (61.6, SD 16.7 y). Among visits

in which mortality occurred, the most common primary admission reasons were brain hemorrhage (256/4320, 5.93%) and sepsis (201/4320, 4.65%). Cardiac arrest and hypoxia showed the highest length of stay, with mean values of 5.3 (SD 5.8) and 5.1 (SD 5.6) days, respectively. For prior Clinical Classifications Software diagnostic history, patients with pneumonia, respiratory failure, arrest, and insufficiency, and shock had a mean length of stay of approximately 7 days (SD 8.6, 8.2, and 8.2, respectively). The proportion of in-hospital mortality events was highest for patients diagnosed with respiratory failure, arrest, and insufficiency, fluid and metabolic disorders, and renal failure ([Table 1](#)).

Table 1. Medical Information Mart for Intensive Care-III patient cohort summary by mortality and average length of stay (N=41,121).

	In-hospital mortality		Length of stay (days)	
	0	1	Values, mean (SD)	Values, median (IQR)
Sex, n (%)				
Male	20,717 (56.29)	2279 (52.75)	3.64 (5.26)	1.99 (1.16-3.74)
Female	16,084 (43.71)	2041 (47.25)	3.63 (5.17)	2.05 (1.17-3.78)
Race, n (%)				
Asian	847 (2.3)	116 (2.69)	3.61 (5.36)	2.0 (1.19-3.54)
Black	3663 (9.95)	286 (6.62)	3.42 (5.01)	1.99 (1.16-3.41)
Hispanic	1369 (3.72)	82 (1.9)	3.24 (4.21)	1.89 (1.12-3.36)
White	26,421 (71.79)	3036 (70.28)	3.61 (5.21)	2.01 (1.16-3.72)
Other or unknown	4501 (12.23)	800 (18.52)	4.04 (5.61)	2.13 (1.2-4.14)
Age				
Not applicable, %	1739 (4.73)	447 (10.35)	— ^a	—
Values, mean (SD)	61.63 (16.71)	68.67 (14.99)	—	—
Height				
Values, mean (SD)	169.01 (13.39)	167.46 (13.22)	—	—
Not applicable, %	28,383 (77.13)	3420 (79.17)	—	—
Weight				
Values, mean (SD)	82.2 (24.14)	77.39 (23.4)	—	—
Not applicable, %	6541 (17.77)	709 (16.41)	—	—
Admission reason, n (%)				
Brain hemorrhage	517 (1.4)	256 (5.93)	4.01 (5.13)	1.93 (1.05-4.74)
Cardiac arrest	105 (0.29)	114 (2.64)	5.27 (5.75)	3.8 (1.7-6.81)
Sepsis	747 (2.03)	201 (4.65)	4.71 (7.19)	2.49 (1.5-4.7)
Respiratory distress	96 (0.26)	25 (0.58)	4.28 (4.45)	2.62 (1.41-5.08)
Liver failure	107 (0.29)	39 (0.9)	4.76 (5.38)	2.78 (1.63-5.8)
Hypoxia	81 (0.22)	21 (0.49)	5.07 (5.61)	2.87 (1.65-5.65)
Cerebrovascular accident	54 (0.15)	18 (0.42)	3.26 (4.29)	1.99 (1.06-3.34)
CCS^b diagnoses: cardiovascular, n (%)				
Acute myocardial infarction	3708 (10.08)	559 (12.94)	3.83 (4.82)	2.25 (1.3-4.2)
Coronary atherosclerosis and other heart disease	12,193 (33.13)	1091 (25.25)	3.22 (4.07)	2.03 (1.17-3.45)
Cardiac dysrhythmias	11,486 (31.21)	1724 (39.91)	4.26 (5.92)	2.3 (1.29-4.44)
Essential hypertension	15,685 (42.62)	1604 (37.13)	3.31 (4.39)	1.99 (1.16-3.43)
Hypertension with complications and secondary hypertension	4738 (12.87)	649 (15.02)	3.77 (5.33)	2.11 (1.22-3.98)
Congestive heart failure and nonhypertensive	9482 (25.77)	1482 (34.31)	4.45 (6.01)	2.54 (1.4-4.85)
Conduction disorders	2603 (7.07)	335 (7.75)	3.58 (4.60)	2.16 (1.24-4.02)
CCS diagnoses: diabetes and metabolic, n (%)				
Fluid and electrolyte disorders	9195 (24.99)	1823 (42.2)	4.51 (5.91)	2.45 (1.38-4.92)
Disorders of lipid metabolism	11,162 (30.33)	812 (18.8)	2.96 (3.77)	1.89 (1.13-3.14)
Diabetes mellitus without complication	7044 (19.14)	872 (20.19)	3.71 (5.12)	2.08 (1.19-3.95)
Diabetes mellitus with complications	3597 (9.77)	288 (6.67)	3.46 (4.96)	2.07 (1.21-3.44)

	In-hospital mortality		Length of stay (days)	
	0	1	Values, mean (SD)	Values, median (IQR)
CCS diagnoses: infectious disease, n (%)				
Septicemia (except in labor)	4332 (11.77)	1497 (34.65)	6.48 (8.53)	3.18 (1.78-7.54)
Pneumonia (except that caused by tuberculosis or sexually transmitted disease)	4594 (12.48)	1094 (25.32)	7.15 (8.60)	3.89 (1.89-8.99)
CCS diagnoses: kidney and gastrointestinal and liver, n (%)				
Acute and unspecified renal failure	6911 (18.78)	1855 (42.94)	5.26 (7.22)	2.77 (1.56-5.75)
Other liver diseases	2847 (7.74)	804 (18.61)	4.98 (6.88)	2.57 (1.44-5.14)
Gastrointestinal hemorrhage	2567 (6.98)	430 (9.95)	4.06 (6.14)	2.09 (1.25-4.04)
Chronic kidney disease	4780 (12.99)	657 (15.21)	3.63 (4.93)	2.12 (1.22-3.91)
CCS diagnoses: respiratory, n (%)				
Respiratory failure, insufficiency, and arrest (adult)	5361 (14.57)	2042 (47.27)	7.30 (8.21)	4.37 (2.1-9.31)
Other upper respiratory disease	1514 (4.11)	143 (3.31)	5.46 (7.31)	2.9 (1.47-6.17)
Other lower respiratory disease	1892 (5.14)	245 (5.67)	4.20 (5.73)	2.25 (1.28-4.56)
Chronic obstructive pulmonary disease and bronchiectasis	4642 (12.61)	684 (15.83)	4.24 (5.86)	2.26 (1.27-4.61)
Pleurisy, pneumothorax, and pulmonary collapse	3139 (8.53)	454 (10.51)	5.78 (7.7)	3.03 (1.7-6.3)
CCS diagnoses: stroke, n (%)				
Acute cerebrovascular disease	2248 (6.11)	798 (18.47)	5.11 (6.51)	2.62 (1.35-6.16)
CCS diagnoses: surgical complications or shock, n (%)				
Complications of surgical procedures or medical care	7823 (21.26)	719 (16.64)	5.18 (7.33)	2.61 (1.36-5.24)
Shock	2017 (5.48)	1173 (27.15)	6.85 (8.15)	3.89 (1.99-8.4)

^aStratified summary statistics were indeterminable for continuous demographic variables.

^bCCS: Clinical Classifications Software.

In-Hospital Mortality Prediction

In a comparison of cross-validation approaches for in-hospital mortality prediction (including all model selection steps and setting the number of folds to 5 for each method), stratified K-fold cross-validation performed approximately the same as regular K-fold cross-validation. Repeated methods of cross-validation performed marginally worse than the simple methods of cross-validation, whereas wider spreads of performance metrics were observed for repeated methods. Nested cross-validation performed slightly worse than both repeated and simpler methods, with a mean AUPR value of 0.369 (compared with 0.371-0.372) and an AUROC value of 0.814 (compared with 0.818-0.821). Across all cross-validation methods, discrimination was moderate to strong for in-hospital mortality prediction, likely owing to the case prevalence of 10.51% (4320/41,121) and the availability of relevant predictive features (demographics, diagnostic history, and admission criteria; [Figure 3](#)).

To assess whether nested cross-validation mitigates overfitting and optimistic bias compared with nonnested methods, we compared the performance estimate given by cross-validation (the average over test folds) with the performance of a model trained on the entire data set with the optimal hyperparameters from cross-validation. We used this refitted model and made

predictions on an entirely withheld validation set (comprising 8224/41,121, 20% of the total data set vs 32,897/41,121, 80% used for cross-validation with model selection). The y-axes in [Figures 4](#) and [5](#) show that the cross-validation estimate had a slight pessimistic bias, as the ratio of validation set performance divided by the cross-validation estimate was >1 . Cross-validation estimates slightly underestimated out-of-sample performance. The discrepancy between the cross-validation estimates and validation set performance was greatest for lower numbers of folds. We only observed marginal differences in the degree of pessimistic bias across the cross-validation methods, although AUPR estimates had a greater bias than AUROC. Estimates from nested cross-validation and repeated K-fold cross-validation were the most pessimistically biased (approximately 1%-2% for AUROC and 5%-9% for AUPR), whereas K-fold cross-validation was the least pessimistically biased ([Figures 4](#) and [5](#)).

Over 100 bootstrap iterations, the 0.632 bootstrap method had a mean AUPR of 0.368 (95% CI 0.351-0.382) and a mean AUROC of 0.819 (95% CI 0.813-0.825). The out-of-bag bootstrap method had a mean AUPR of 0.367 (95% CI 0.346-0.390) and a mean AUROC of 0.818 (95% CI 0.796-0.828).

We also repeated the optimism estimation experiment using cross-validation methods (each specified with 5 folds) applied

to 10 randomly sampled validation sets for a more robust estimation of the model performance bias. Nested cross-validation showed a marginally greater pessimistic bias than nonnested cross-validation methods for both AUROC and AUPR. Over the 10 randomly sampled validation sets, outlier values for the relative error of the cross-validation estimate ranged from 8% optimistic bias (AUPR for nested cross-validation) to 10% pessimistic bias (AUPR for all nonnested methods; Figures 6 and 7).

In addition to modest performance differences, tendencies toward increased computational time were observed with a more sophisticated schema, for example, nested cross-validation. Although the overall training time differences were inconsequential for this data set, the computational time of the nested cross-validation increased quadratically with the number of folds ($O(k^2)$). In comparison, the computational time required for the repeated cross-validation methods increased linearly ($O(k)$) and simple cross-validation methods required nearly constant time across varying number of folds ($O(c)$) (Figure 8).

Figure 3. Discrimination for in-hospital mortality prediction by cross-validation (CV) method (with 5 folds used for each method). AUPR: area under the precision-recall curve; AUROC: area under the receiver operator characteristic curve.

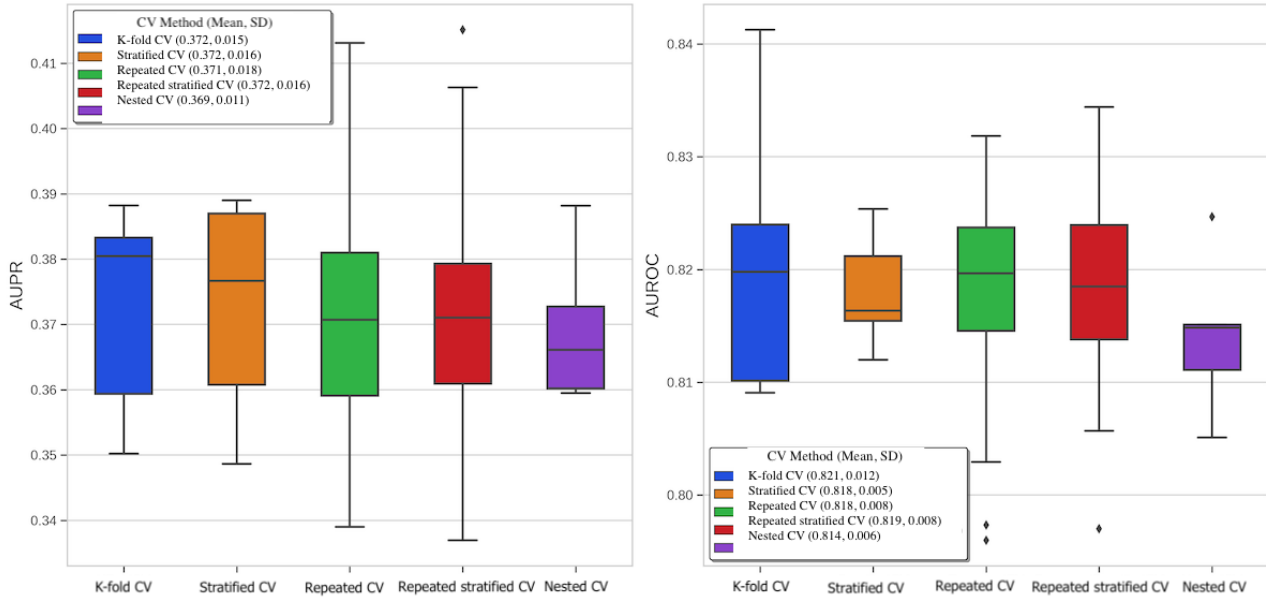


Figure 4. Cross-validation (CV) estimates versus validation set performance (AUROC) for in-hospital mortality by number of folds. AUROC: area under the receiver operator characteristic curve.

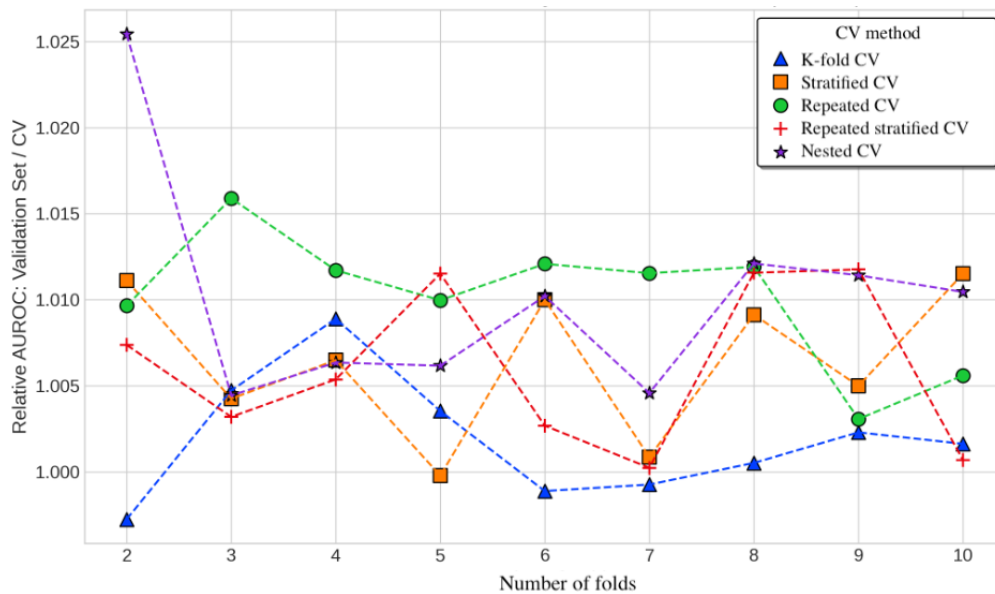


Figure 5. Cross-validation (CV) estimates versus validation set performance (AUPR) for in-hospital mortality by number of folds. AUPR: area under the precision-recall curve.

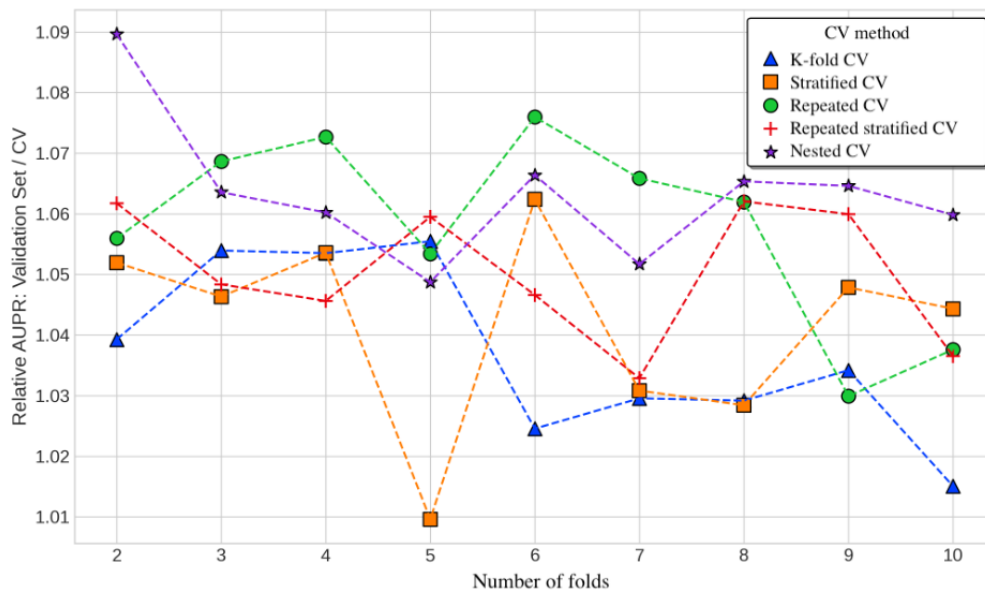


Figure 6. Cross-validation (CV) estimates versus validation set performance (AUROC) for in-hospital mortality over repeated 5-fold trials. AUROC: area under the receiver operator characteristic curve.

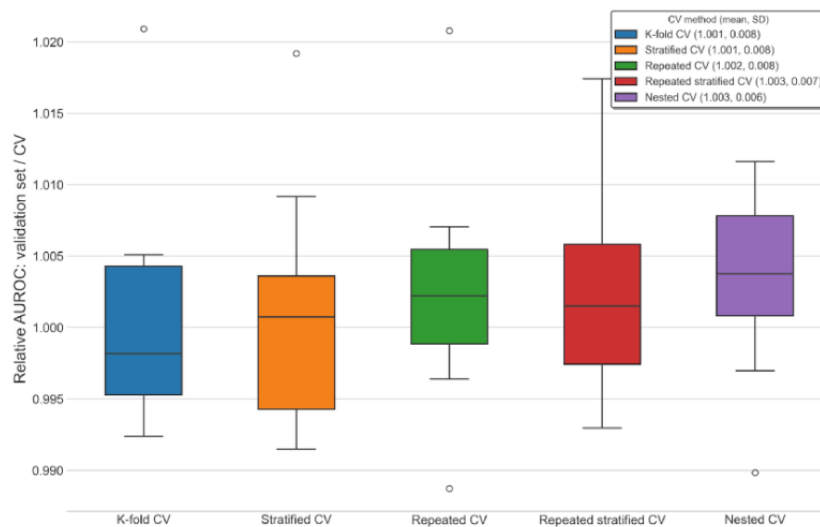


Figure 7. Cross-validation (CV) estimates versus validation set performance (AUPR) for in-hospital mortality over repeated 5-fold trials. AUPR: area under the precision-recall curve.

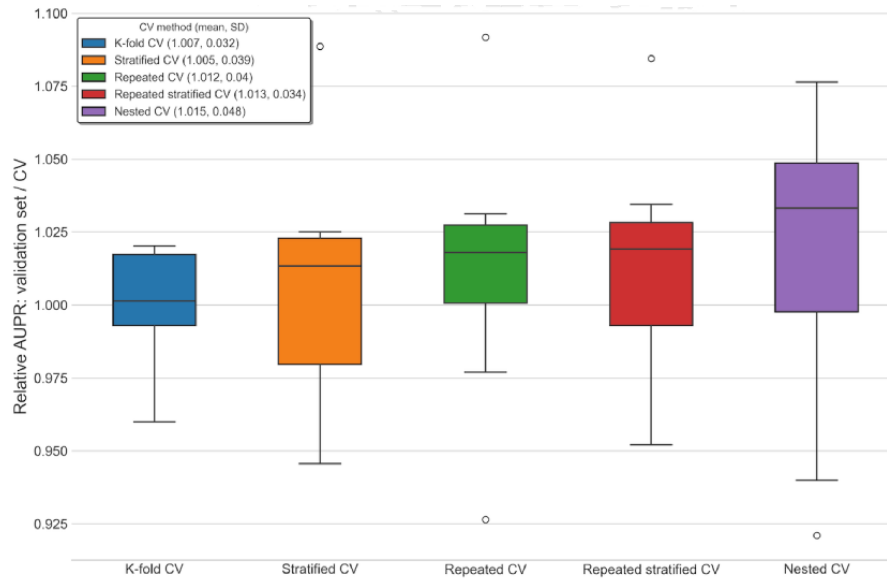
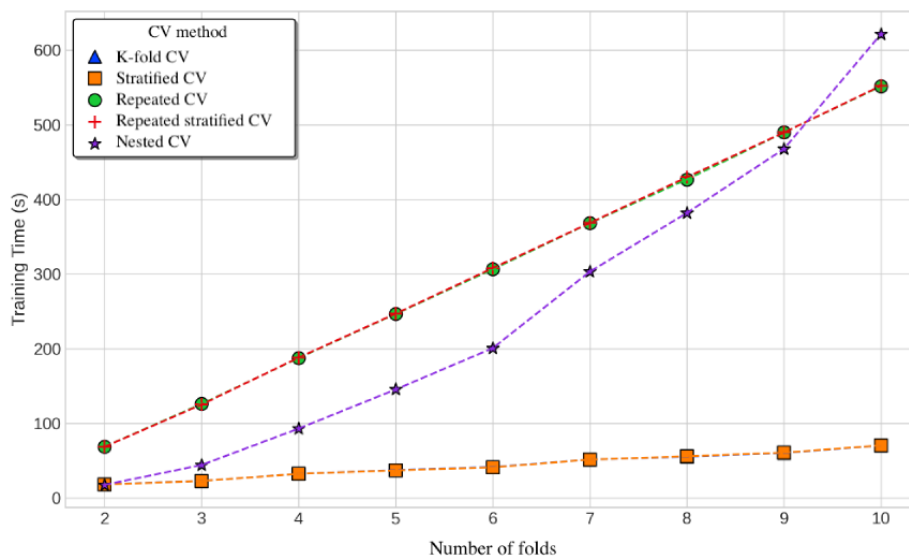


Figure 8. Computational time required by cross-validation (CV) method and number of folds for in-hospital mortality prediction. CV: cross-validation.



Length of Stay Prediction

With length of stay prediction defined as a regression problem, we compared the test-fold performance metrics across various cross-validation methods (with each method using 5 folds). We were unable to include stratified cross-validation, which is only applicable to classification problems wherein the case prevalence can be made equivalent across different training and test folds. Similar to in-hospital mortality prediction, we observed equivalent or marginally worse performance for nested cross-validation compared with nonnested methods (with average mean absolute errors of 2.39 vs 2.38 for nested vs nonnested methods, and average median absolute errors of 1.23 for all methods). The mean absolute error was nearly twice that of the median absolute error, which suggests that higher outlier values for length of stay increased the mean prediction error in this regression problem (Figure 9).

The 0.632 bootstrap method had an average mean absolute error of 2.01 (95% CI 1.98-2.04) and an average median absolute error of 1.05 (95% CI 1.03-1.07). The out-of-bag bootstrap method had an average mean absolute error of 2.84 (95% CI 2.79-2.90) and an average median absolute error of 1.53 (95% CI 1.49-1.55).

For median absolute error, all cross-validation methods showed a slight pessimistic bias (because the validation set performance was slightly greater than the estimated performance from cross-validation). There were few disparities between the accuracy of the performance estimates for varying numbers of folds or different cross-validation methods. The pessimistic bias was greatest for the K-fold cross-validation with 2 folds (approximately 2%). Nested cross-validation produced the least biased estimates overall, although the bias from the nonnested methods remained <1% (Figure 10).

Repeated and single K-fold cross-validation estimates of the mean absolute error were slightly optimistically biased, whereas

nested cross-validation was pessimistically biased across all folds. As the number of folds increased, the nonnested methods were generally less biased. Nested and nonnested cross-validation estimates of the mean absolute error were approximately unbiased, with a range between 1% pessimistic bias (nested cross-validation) and 1% optimistic bias (K-fold cross-validation; Figure 11).

Consistent with the classification problem, we also observed a quadratic relationship between the computational time required for nested cross-validation and the number of folds. K-fold cross-validation showed linear time complexity, with repeated K-fold cross-validation increasing linearly with the additional multiplicative factor from the number of repeats. Owing to the

increased training time required for ensemble models such as random forest, the absolute time required for cross-validation methods was much higher for length of stay prediction than for in-hospital mortality (Figure 12).

Finally, we tested record-wise versus subject-wise cross-validation and found negligible differences in the accuracy of the model performance estimates. We suspect that this may have resulted from the relatively few repeated hospital visits (records or observations in our data set) associated with each unique subject or the minimal correlation between the feature values of identical records across different hospital visits (ie, differences in reasons for hospital admission may have been diverse within a subject's set of visit records).

Figure 9. Regression metrics by cross-validation (CV) method for length of stay prediction (with 5 folds used for each method).

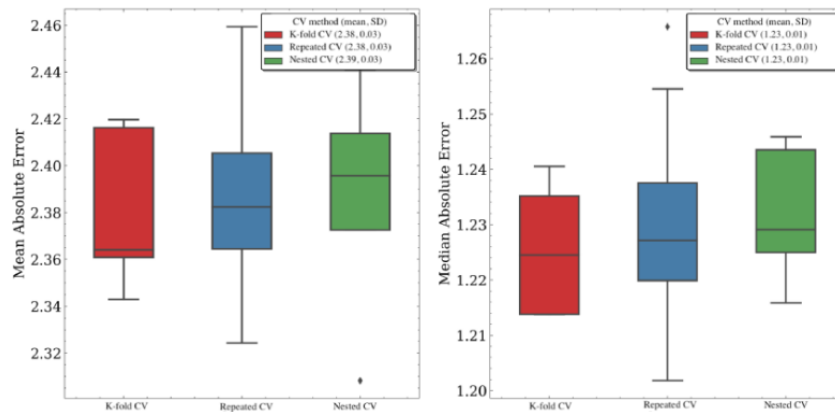


Figure 10. Cross-validation (CV) estimates versus validation set performance (Median Absolute Error) by number of folds for length of stay prediction.

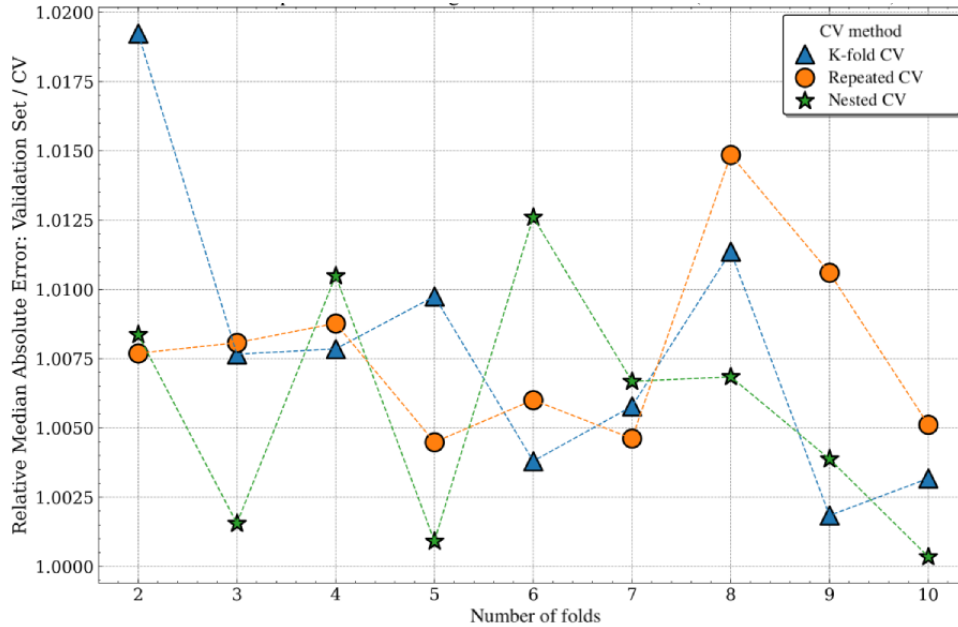


Figure 11. Cross-validation (CV) estimates versus validation set performance (Mean Absolute Error) by number of folds for length of stay prediction.

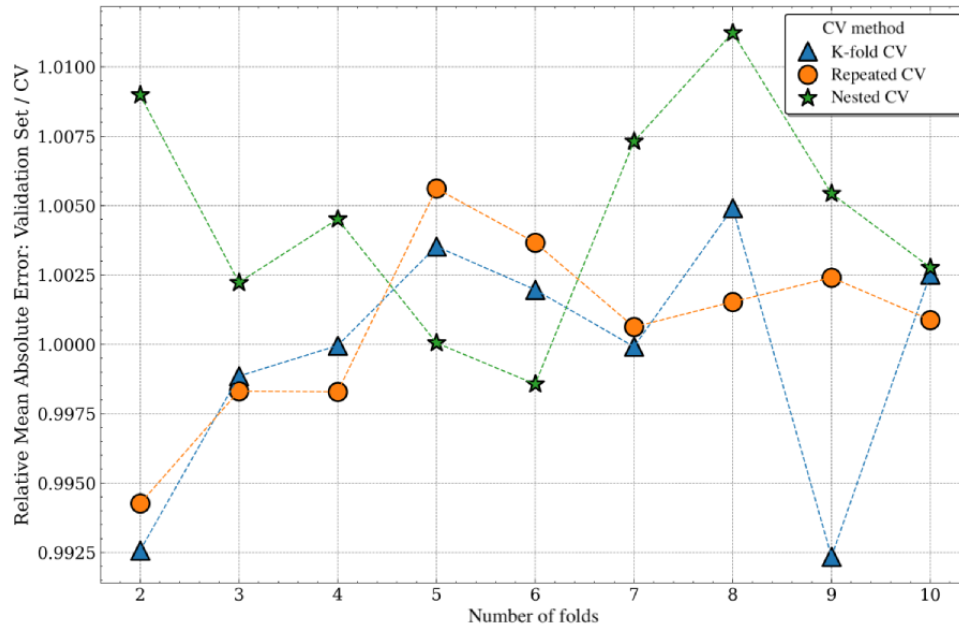
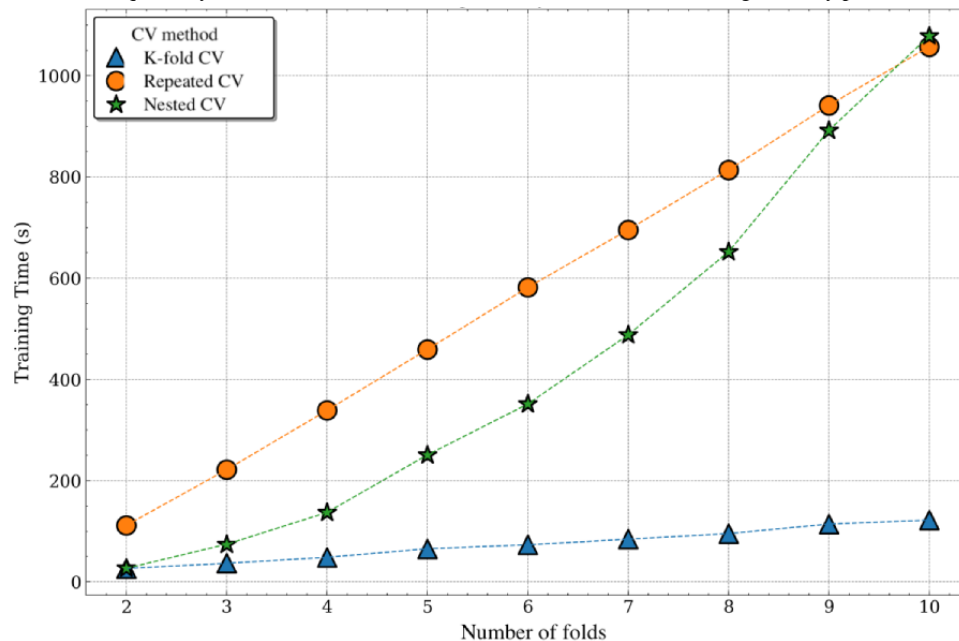


Figure 12. Computational time required by cross-validation (CV) method and number of folds for length of stay prediction. CV: cross-validation.



Recommendations, Common Missteps, and Best Practices

This tutorial described and compared multiple forms of cross-validation. Cross-validation generally results in reduced bias compared with holdout testing and poses the clear advantage of training and testing on all available data [6]. A more sophisticated schema of model validation involves bootstrapping methods (and even involving bootstrap-based cross-validation or repeated nested cross-validation). However, the modest computational time and the acceptable biased estimates of true test error that we observed suggest that the conventional cross-validation methods should remain the first line for real-world health care modeling. Although K-fold cross-validation remains the most common, other types of

cross-validation pose advantages and disadvantages worth considering for each use case (Multimedia Appendix 1). Case studies using readily accessible and well-studied EHR data, MIMIC-III, showed slight performance differences in terms of cross-validation performance and optimistic bias for more computationally intensive forms of cross-validation such as nested cross-validation. Although our results should not dictate whether nested cross-validation is used across the variety of prediction problems and clinical data sets, the reduction in optimistic bias with nested cross-validation does not outweigh the additional challenges of implementing nested cross-validation and the added computational time it requires.

Common missteps detract from the potential for cross-validation in diverse modeling scenarios. For example, model development might be more complex across iterations, and separating

development (feature selection, hyperparameter tuning, and classifier selection) from model validation remains paramount. When using cross-validation for model development within the same process as model validation, it is often recommended to use a nested cross-validation approach where preprocessing, feature selection, and hyperparameter tuning are conducted entirely independent of the “outer fold” that is used independently for validation.

However, as our results illustrate, and prior studies have covered both empirically and theoretically, the degree of optimistic bias from nonnested cross-validation methods varies with the number of features, the size of the data set, the extent of hyperparameter tuning and feature selection, and the nature of the clinical outcome. In this case study, we had a relatively large sample size relative to the number of original features included (referred to as $n > p$). When performing hyperparameter tuning and other model selection steps within cross-validation, the difference in optimistic bias observed between nested and nonnested cross-validation methods should have been mitigated [31]. Furthermore, our range of hyperparameters and the subsequent number of possible model tuning configurations was relatively smaller than a developer would typically use when wanting to optimize performance. This also contributed to a relatively lower reduction in optimistic bias when using the theoretically validated approach to reduce optimistic bias resulting from model tuning in cross-validation (nested methods) [12,30,32].

Although it is impossible to recommend a single cross-validation approach that will be appropriate for all modeling scenarios, we encourage developers to use nested cross-validation methods in cases with higher dimensional feature spaces relative to the sample size, higher numbers of algorithms and parameters being tested, and problems in which the increase in computational time required from nested cross-validation remains within feasible bounds (as we observed in both modeling problems in our case study). As emphasized throughout our discussion of nested cross-validation, this approach also offers the simplicity of performing both model selection and tuning and model evaluation within the same procedure, allowing developers to disregard concerns about or additional evaluations needed to mitigate the bias introduced when using nonnested methods for model selection (which should be used by default to optimize model performance) and model evaluation. Although we hope to contribute further empirical evidence on the comparative bias between various cross-validation methods, we emphasize that this work is a tutorial meant to demonstrate the use of various approaches that developers can use for their specific use cases.

In routine health care, EHRs include repeated, irregular samples (records or health care encounters) across records (patients). Although we observed negligible differences in performance between subject-wise and record-wise cross-validation in this case study, the use case for predictive modeling should determine the choice between subject-wise and record-wise sampling. For example, a cohort study of encounters in an emergency department to predict admissions for pneumonia might include data sets with multiple encounters per person, some with a single encounter and others with multiple encounters. Record-wise splitting might permit encounters for the same individual to be present in both training and testing sets, even if the outcomes of each of those encounters with respect to the prediction target might differ. The tendency in health care data for correlation and, specifically, autocorrelation would also introduce undue bias in this scenario.

A fundamental misconception about cross-validation is that it necessarily “returns” a model that can then be used for production deployment or external validation [33]. Rather, cross-validation is more appropriately considered a *learning procedure*, which allows a developer to fine-tune the parameters involved in model development and estimate model performance on out-of-sample data (internal validation). Once model selection via cross-validation has produced the best selected features, hyperparameters, and modeling algorithm, it is necessary to retrain a “final model” using the entire available data set with these optimized specifications.

We hope to address the current limitations of machine learning evaluation and development that might hinder the translation and reproducibility of predictive models in health care. With respect to their specific clinical implications, we provided greater conceptual understanding of cross-validation as both a model evaluation and model development method, outlined the respective strengths and weaknesses of common cross-validation methods, specified the technical steps involved when using cross-validation with model tuning and selection, demonstrated cross-validation in a real-world case study, and offered further empirical evidence on the performance and computational time of cross-validation methods. Practically, we refer readers to our open-source code repository with reproducible Jupyter notebooks and Python code, implementing all the statistical analyses and experiments of this tutorial. Therefore, developers will have access to cross-validation examples with real-world health care data and software functionality that can aid developers with various clinical machine learning problems.

Acknowledgments

CGW receives support from the National Institute of Mental Health (R01MH118233, R01MH121455, R01MH116269, R01MH120122), the National Human Genome Research Institute (RM1HG009034), the National Institutes of Health (U54HG012510), FDA Sentinel (WO2006), and Wellcome Leap MCPsych.

Conflicts of Interest

None declared.

Multimedia Appendix 1

<https://ai.jmir.org/2023/1/e49023>

JMIR AI 2023 | vol. 2 | e49023 | p.23
(page number not for citation purposes)

Major types of cross-validation with broad advantages and disadvantages.

[[DOCX File , 27 KB - ai_v2i1e49023_app1.docx](#)]

References

1. Lindsell CJ, Stead WW, Johnson KB. Action-informed artificial intelligence-matching the algorithm to the problem. *JAMA* 2020 Jun 02;323(21):2141-2142. [doi: [10.1001/jama.2020.5035](https://doi.org/10.1001/jama.2020.5035)] [Medline: [32356878](https://pubmed.ncbi.nlm.nih.gov/32356878/)]
2. Steyerberg EW, Vergouwe Y. Towards better clinical prediction models: seven steps for development and an ABCD for validation. *Eur Heart J* 2014 Aug 01;35(29):1925-1931 [FREE Full text] [doi: [10.1093/eurheartj/ehu207](https://doi.org/10.1093/eurheartj/ehu207)] [Medline: [24898551](https://pubmed.ncbi.nlm.nih.gov/24898551/)]
3. Bedoya AD, Economou-Zavlanos NJ, Goldstein BA, Young A, Jelovsek JE, O'Brien C, et al. A framework for the oversight and local deployment of safe and high-quality prediction models. *J Am Med Inform Assoc* 2022 Aug 16;29(9):1631-1636 [FREE Full text] [doi: [10.1093/jamia/ocac078](https://doi.org/10.1093/jamia/ocac078)] [Medline: [35641123](https://pubmed.ncbi.nlm.nih.gov/35641123/)]
4. Bouwmeester W, Zuithoff NP, Mallett S, Geerlings MI, Vergouwe Y, Steyerberg EW, et al. Reporting and methods in clinical prediction research: a systematic review. *PLoS Med* 2012 May 22;9(5):1-12 [FREE Full text] [doi: [10.1371/journal.pmed.1001221](https://doi.org/10.1371/journal.pmed.1001221)] [Medline: [22629234](https://pubmed.ncbi.nlm.nih.gov/22629234/)]
5. Collins GS, Omar O, Shanyinde M, Yu LM. A systematic review finds prediction models for chronic kidney disease were poorly reported and often developed using inappropriate methods. *J Clin Epidemiol* 2013 Mar;66(3):268-277. [doi: [10.1016/j.jclinepi.2012.06.020](https://doi.org/10.1016/j.jclinepi.2012.06.020)] [Medline: [23116690](https://pubmed.ncbi.nlm.nih.gov/23116690/)]
6. Smith GC, Seaman SR, Wood AM, Royston P, White IR. Correcting for optimistic prediction in small data sets. *Am J Epidemiol* 2014 Aug 01;180(3):318-324 [FREE Full text] [doi: [10.1093/aje/kwu140](https://doi.org/10.1093/aje/kwu140)] [Medline: [24966219](https://pubmed.ncbi.nlm.nih.gov/24966219/)]
7. Moons KG, Kengne AP, Woodward M, Royston P, Vergouwe Y, Altman DG, et al. Risk prediction models: I. Development, internal validation, and assessing the incremental value of a new (bio)marker. *Heart* 2012 May 07;98(9):683-690. [doi: [10.1136/heartjnl-2011-301246](https://doi.org/10.1136/heartjnl-2011-301246)] [Medline: [22397945](https://pubmed.ncbi.nlm.nih.gov/22397945/)]
8. Moons KG, Kengne AP, Grobbee DE, Royston P, Vergouwe Y, Altman DG, et al. Risk prediction models: II. External validation, model updating, and impact assessment. *Heart* 2012 May 07;98(9):691-698. [doi: [10.1136/heartjnl-2011-301247](https://doi.org/10.1136/heartjnl-2011-301247)] [Medline: [22397946](https://pubmed.ncbi.nlm.nih.gov/22397946/)]
9. Steyerberg EW, Harrell FEJ. Prediction models need appropriate internal, internal-external, and external validation. *J Clin Epidemiol* 2016 Jan;69:245-247 [FREE Full text] [doi: [10.1016/j.jclinepi.2015.04.005](https://doi.org/10.1016/j.jclinepi.2015.04.005)] [Medline: [25981519](https://pubmed.ncbi.nlm.nih.gov/25981519/)]
10. Tougui I, Jilbab A, Mhamdi JE. Impact of the choice of cross-validation techniques on the results of machine learning-based diagnostic applications. *Healthc Inform Res* 2021 Jul;27(3):189-199 [FREE Full text] [doi: [10.4258/hir.2021.27.3.189](https://doi.org/10.4258/hir.2021.27.3.189)] [Medline: [34384201](https://pubmed.ncbi.nlm.nih.gov/34384201/)]
11. Tohka J, van Gils M. Evaluation of machine learning algorithms for health and wellness applications: a tutorial. *Comput Biol Med* 2021 May;132:104324 [FREE Full text] [doi: [10.1016/j.combiomed.2021.104324](https://doi.org/10.1016/j.combiomed.2021.104324)] [Medline: [33774270](https://pubmed.ncbi.nlm.nih.gov/33774270/)]
12. Varoquaux G, Raamana PR, Engemann DA, Hoyos-Idrobo A, Schwartz Y, Thirion B. Assessing and tuning brain decoders: cross-validation, caveats, and guidelines. *Neuroimage* 2017 Jan 15;145(Pt B):166-179. [doi: [10.1016/j.neuroimage.2016.10.038](https://doi.org/10.1016/j.neuroimage.2016.10.038)] [Medline: [27989847](https://pubmed.ncbi.nlm.nih.gov/27989847/)]
13. Ambroise C, McLachlan GJ. Selection bias in gene extraction on the basis of microarray gene-expression data. *Proc Natl Acad Sci U S A* 2002 May 14;99(10):6562-6566 [FREE Full text] [doi: [10.1073/pnas.102102699](https://doi.org/10.1073/pnas.102102699)] [Medline: [11983868](https://pubmed.ncbi.nlm.nih.gov/11983868/)]
14. Johnson A, Bulgarelli L, Pollard T, Horng S, Celi LA, Mark R. MIMIC-IV Clinical database demo (version 1.0). *PhysioNet*. 2022. URL: <https://doi.org/10.13026/jwtp-v091> [accessed 2023-11-03]
15. Cross-Validation Tutorial: Predictive Modeling in Healthcare. github. URL: https://github.com/drewwilimitis/JMIR_CV_Tutorial [WebCite Cache ID https://github.com/drewwilimitis/JMIR_CV_Tutorial]
16. Larson SC. The shrinkage of the coefficient of multiple correlation. *J Educ Psychol* 1931 Jan;22(1):45-55. [doi: [10.1037/h0072400](https://doi.org/10.1037/h0072400)]
17. Mosteller F, Tukey JW. Data analysis, including statistics. In: Lindzey G, Aronson E, editors. *Handbook of Social Psychology*, Volume 2. Boston, MA: Addison-Wesley; 1968:1-26.
18. Little MA, Varoquaux G, Saeb S, Lonini L, Jayaraman A, Mohr DC, et al. Using and understanding cross-validation strategies. *Perspectives on Saeb et al. Gigascience* 2017 May 01;6(5):1-6 [FREE Full text] [doi: [10.1093/gigascience/gix020](https://doi.org/10.1093/gigascience/gix020)] [Medline: [28327989](https://pubmed.ncbi.nlm.nih.gov/28327989/)]
19. Saeb S, Lonini L, Jayaraman A, Mohr DC, Kording KP. The need to approximate the use-case in clinical machine learning. *Gigascience* 2017 May 01;6(5):1-9 [FREE Full text] [doi: [10.1093/gigascience/gix019](https://doi.org/10.1093/gigascience/gix019)] [Medline: [28327985](https://pubmed.ncbi.nlm.nih.gov/28327985/)]
20. 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, Quebec URL: <https://www.ijcai.org/Proceedings/95-2/Papers/016.pdf>
21. Steyerberg EW, Vickers AJ, Cook NR, Gerds T, Gonen M, Obuchowski N, et al. Assessing the performance of prediction models: a framework for traditional and novel measures. *Epidemiology* 2010 Jan;21(1):128-138 [FREE Full text] [doi: [10.1097/EDE.0b013e3181c30fb2](https://doi.org/10.1097/EDE.0b013e3181c30fb2)] [Medline: [20010215](https://pubmed.ncbi.nlm.nih.gov/20010215/)]
22. Steyerberg EW. *Clinical Prediction Models: A Practical Approach to Development, Validation, and Updating*. New York, NY: Springer; 2009.

23. Efron B. Estimating the error rate of a prediction rule: improvement on cross-validation. *J Am Stat Assoc* 1983 Jun;78(382):316-331. [doi: [10.1080/01621459.1983.10477973](https://doi.org/10.1080/01621459.1983.10477973)]
24. Efron B, Tibshirani R. Improvements on cross-validation: the 632+ bootstrap method. *J Am Stat Assoc* 1997 Jun;92(438):548-560. [doi: [10.1080/01621459.1997.10474007](https://doi.org/10.1080/01621459.1997.10474007)]
25. Molinaro AM, Simon R, Pfeiffer RM. Prediction error estimation: a comparison of resampling methods. *Bioinformatics* 2005 Aug 01;21(15):3301-3307. [doi: [10.1093/bioinformatics/bti499](https://doi.org/10.1093/bioinformatics/bti499)] [Medline: [15905277](https://pubmed.ncbi.nlm.nih.gov/15905277/)]
26. Krstajic D, Buturovic LJ, Leahy DE, Thomas S. Cross-validation pitfalls when selecting and assessing regression and classification models. *J Cheminform* 2014 Mar 29;6(1):10 [FREE Full text] [doi: [10.1186/1758-2946-6-10](https://doi.org/10.1186/1758-2946-6-10)] [Medline: [24678909](https://pubmed.ncbi.nlm.nih.gov/24678909/)]
27. Li M, Du S. Current status and trends in researches based on public intensive care databases: a scientometric investigation. *Front Public Health* 2022 Sep 15;10:912151 [FREE Full text] [doi: [10.3389/fpubh.2022.912151](https://doi.org/10.3389/fpubh.2022.912151)] [Medline: [36187634](https://pubmed.ncbi.nlm.nih.gov/36187634/)]
28. Stone M. Cross - validity choice and assessment of statistical predictions. *J R Stat Soc Series B Stat Methodol* 2018 Dec 05;36(2):111-133. [doi: [10.1111/j.2517-6161.1974.tb00994.x](https://doi.org/10.1111/j.2517-6161.1974.tb00994.x)]
29. Varma S, Simon R. Bias in error estimation when using cross-validation for model selection. *BMC Bioinformatics* 2006 Feb 23;7(1):91 [FREE Full text] [doi: [10.1186/1471-2105-7-91](https://doi.org/10.1186/1471-2105-7-91)] [Medline: [16504092](https://pubmed.ncbi.nlm.nih.gov/16504092/)]
30. Cawley GC, Talbot NL. On over-fitting in model selection and subsequent selection bias in performance evaluation. *J Mach Learn Res* 2010;11:2079-2107.
31. Bates S, Hastie T, Tibshirani R. Cross-validation: what does it estimate and how well does it do it? *J Am Stat Assoc* 2023 May 15:1-12 [FREE Full text] [doi: [10.1080/01621459.2023.2197686](https://doi.org/10.1080/01621459.2023.2197686)]
32. Rao RB, Fung G. On the dangers of cross-validation. An experimental evaluation. In: *Proceedings of the SIAM International Conference on Data Mining*. 2008 Presented at: SIAM International Conference on Data Mining; April 24-26, 2008; Atlanta, GA. [doi: [10.1137/1.9781611972788.54](https://doi.org/10.1137/1.9781611972788.54)]
33. Tsamardinos I, Greasidou E, Borboudakis G. Bootstrapping the out-of-sample predictions for efficient and accurate cross-validation. *Mach Learn* 2018 May 9;107(12):1895-1922 [FREE Full text] [doi: [10.1007/s10994-018-5714-4](https://doi.org/10.1007/s10994-018-5714-4)] [Medline: [30393425](https://pubmed.ncbi.nlm.nih.gov/30393425/)]

Abbreviations

AUPR: area under the precision-recall curve

AUROC: area under the receiver operator characteristic curve

EHR: electronic health record

MIMIC-III: Medical Information Mart for Intensive Care-III

Edited by B Malin, K El Emam; submitted 15.05.23; peer-reviewed by U Sinha, S Figini; comments to author 05.09.23; revised version received 19.09.23; accepted 28.09.23; published 18.12.23.

Please cite as:

Wilimitis D, Walsh CG

Practical Considerations and Applied Examples of Cross-Validation for Model Development and Evaluation in Health Care: Tutorial
JMIR AI 2023;2:e49023

URL: <https://ai.jmir.org/2023/1/e49023>

doi: [10.2196/49023](https://doi.org/10.2196/49023)

PMID: [38875530](https://pubmed.ncbi.nlm.nih.gov/38875530/)

©Drew Wilimitis, Colin G Walsh. Originally published in JMIR AI (<https://ai.jmir.org>), 18.12.2023. This is an open-access article distributed under the terms of the Creative Commons Attribution License (<https://creativecommons.org/licenses/by/4.0/>), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR AI, is properly cited. The complete bibliographic information, a link to the original publication on <https://www.ai.jmir.org/>, as well as this copyright and license information must be included.

Review

Strategies to Improve the Impact of Artificial Intelligence on Health Equity: Scoping Review

Carl Thomas Berdahl^{1,2,3}, MD, MS; Lawrence Baker¹, MSc; Sean Mann¹, MSc; Osonde Osoba¹, MSc, PhD; Federico Girosi¹, PhD

¹RAND Corporation, Santa Monica, CA, United States

²Department of Medicine, Cedars-Sinai Medical Center, Los Angeles, CA, United States

³Department of Emergency Medicine, Cedars-Sinai Medical Center, Los Angeles, CA, United States

Corresponding Author:

Carl Thomas Berdahl, MD, MS

RAND Corporation

1776 Main Street

Santa Monica, CA, 90401

United States

Phone: 1 3104233091

Email: cberdahl@rand.org

Abstract

Background: Emerging artificial intelligence (AI) applications have the potential to improve health, but they may also perpetuate or exacerbate inequities.

Objective: This review aims to provide a comprehensive overview of the health equity issues related to the use of AI applications and identify strategies proposed to address them.

Methods: We searched PubMed, Web of Science, the IEEE (Institute of Electrical and Electronics Engineers) Xplore Digital Library, ProQuest U.S. Newsstream, Academic Search Complete, the Food and Drug Administration (FDA) website, and ClinicalTrials.gov to identify academic and gray literature related to AI and health equity that were published between 2014 and 2021 and additional literature related to AI and health equity during the COVID-19 pandemic from 2020 and 2021. Literature was eligible for inclusion in our review if it identified at least one equity issue and a corresponding strategy to address it. To organize and synthesize equity issues, we adopted a 4-step AI application framework: Background Context, Data Characteristics, Model Design, and Deployment. We then created a many-to-many mapping of the links between issues and strategies.

Results: In 660 documents, we identified 18 equity issues and 15 strategies to address them. Equity issues related to Data Characteristics and Model Design were the most common. The most common strategies recommended to improve equity were improving the quantity and quality of data, evaluating the disparities introduced by an application, increasing model reporting and transparency, involving the broader community in AI application development, and improving governance.

Conclusions: Stakeholders should review our many-to-many mapping of equity issues and strategies when planning, developing, and implementing AI applications in health care so that they can make appropriate plans to ensure equity for populations affected by their products. AI application developers should consider adopting equity-focused checklists, and regulators such as the FDA should consider requiring them. Given that our review was limited to documents published online, developers may have unpublished knowledge of additional issues and strategies that we were unable to identify.

(JMIR AI 2023;2:e42936) doi:[10.2196/42936](https://doi.org/10.2196/42936)

KEYWORDS

artificial intelligence; machine learning; health equity; health care disparities; algorithmic bias; social determinants of health; decision making; algorithms; gray literature; equity; health data

Introduction

Background and Rationale

The use of artificial intelligence (AI) in clinical care and public health contexts has expanded rapidly in recent years [1-6], including throughout the COVID-19 pandemic [7-15]. While emerging AI applications have the potential to improve health care quality and fairness [16-21], they may alternatively perpetuate or exacerbate inequities if they are not designed, deployed, and monitored appropriately [22-26].

Health equity is defined by the World Health Organization as “the absence of unfair and avoidable or remediable differences in health among population groups defined socially, economically, demographically, or geographically.... Pursuing health equity means...giving special attention to the needs of those at greatest risk of poor health, based on social conditions.” [27]. According to the Robert Wood Johnson Foundation, “achieving health equity requires identifying and addressing not only overt discrimination but also unconscious and implicit bias and the discriminatory effects—intended and unintended—of structures and policies created by historical injustices, even when conscious intent is no longer clearly present.” [28].

Concerns about AI’s impact on health equity have been discussed extensively in academic and gray literature. Several frameworks identify AI health equity issues throughout development and propose strategies to address them. For example, Chen et al [29] created a 5-step ethical pipeline for health care model development and recommended best practices at each step. Others have proposed similar 6-, 5-, or 4-step frameworks [21,30,31]. Catering more directly to practitioners, researchers at Chicago Booth created an “algorithmic bias playbook” [32]: step-by-step instructions for organizations to identify, improve, and protect against biased algorithms so that fairness is enhanced for vulnerable populations. These frameworks focus on developers as the stakeholder with both the responsibility and the means to improve health equity outcomes. A recent report from Imperial College London built upon Chen et al’s framework to further describe several health equity issues, suggest more detailed strategies, and advocate for action from a broader range of stakeholders, including policymakers [33].

While the aforesaid frameworks related to AI and equity were disseminated between 2016 and 2022, none link equity strategies to multiple issues. An investigation identifying links between health equity issues and strategies to address them is warranted so that stakeholders can understand the universe of approaches to improve health equity at all stages of AI application development and deployment.

Objectives

The objective of this scoping review was to identify equity issues for health AI applications and connect each issue with corresponding strategies. In addition, we sought to produce a framework that would be useful to independent evaluators whose role is to make comprehensive recommendations for strategies to address equity-relevant issues.

The objective of this review was established in consultation with the study sponsor as part of a broader project examining AI, COVID-19, and health equity. Stakeholder consultation, initial document searches, and document screening were undertaken as part of this broader project and are also described in a separate article on the use of AI in the COVID-19 response [34].

Methods

Overview

We adopted a scoping review approach [35] to identify and describe equity issues arising due to implementation of AI in health and catalog strategies to address each issue. In performing the scoping review, we followed the 5 steps described by Arksey and O’Malley [35], although we opted to begin the recommended optional stakeholder consultation before conducting the literature review so that our stakeholders could assist with our search strategy development. We elected a scoping review approach because it is well-suited to “[summarize] findings from a body of knowledge that is heterogeneous in methods or discipline” such as available academic and gray literature [36]. We followed the PRISMA-ScR (Preferred Reporting Items for Systematic Reviews and Meta-Analyses Extension for Scoping Reviews; [Multimedia Appendix 1](#)) reporting guidelines as we designed and executed our review [36]. While the study protocol is not published online, [Multimedia Appendix 2](#) includes a detailed description of the search strategy.

Preparatory Stakeholder Consultation

To best understand the contextual landscape of our scoping review, we began our project by consulting a diverse group of 9 health care stakeholders: 1 patient advocate, 2 clinicians, 1 health system representative, 1 health insurance representative, 1 public policymaker, 1 public health official, 1 industry representative, and 1 researcher. Interviews with these stakeholders helped us define what was in scope for our review and refine inclusion and exclusion criteria for our literature search strategy. The stakeholders we interviewed also identified exemplar peer-reviewed and gray literature documents, existing frameworks, and example lists of issues and strategies. The stakeholder interview protocol, which was provided to stakeholders and also covered topics related to AI and health equity as part of a broader research study, is available in [Multimedia Appendix 2](#).

Eligibility Criteria

Documents were considered eligible for inclusion in our literature search if (1) they were available in the English language, (2) they related to AI, and (3) they discussed health equity or the clinical or public health response to COVID-19. For documents unrelated to COVID-19, the literature search included publications between January 1, 2014, and December 10, 2021. For documents related to COVID-19, the literature search was limited to the period from December 31, 2019, to December 2021.

Information Sources and Search Strategy

We searched 3 databases to identify academic literature of interest: PubMed, Web of Science, and the IEEE (Institute of Electrical and Electronics Engineers) Xplore Digital Library. As directed by the medical reference librarian who assisted us with our search strategy, we also searched 2 databases to identify news articles and media commentaries of interest, which she believed would be important to identifying emerging issues and

strategies that had not yet been evaluated by academic researchers: ProQuest U.S. Newsstream and Academic Search Complete. Finally, we searched the Food and Drug Administration website and ClinicalTrials.gov for documents meeting inclusion criteria. [Textbox 1](#) gives an overview of our search strategy according to the BeHEMoTh (Behavior, Health condition, Exclusions, Models, or Theories) framework [37]. Detailed parameters for the search strategy are provided in [Multimedia Appendix 2](#).

Textbox 1. Search strategy outline using the BeHEMoTh framework [37].

- *Behavior of interest (artificial intelligence):* artificial intelligence, machine learning, deep learning, supervised learning, unsupervised learning, reinforcement learning unsupervised clustering, unsupervised classification, supervised classification, natural language processing, expert system, rules engine, fuzzy logic, or algorithm.
- *Health context (clinical or public health response to COVID-19):* health, clinic, hospital, therapy, medical, care, COVID-19, public health
- *Model or theory (equity):* equity, fairness, bias, inequality, race, gender, sex, gender, social determinants of health, socioeconomic status, income, minority, disadvantaged, vulnerable, marginalized, disparities, prejudiced, or minority.
- *Exclusions:* documents in a language other than English.

To be included in our review, a document had to relate to the behavior of interest (artificial intelligence) and at least one of the following: the health context (clinical or public health response to COVID-19) or the model or theory (equity).

Selection of Documents and Data Charting Process

We screened all documents of potential interest to determine which were eligible for full-text review. Articles of potential interest were added to a Microsoft Excel spreadsheet to facilitate the selection process and data charting of our progress. If an article did not have an abstract, it was automatically eligible for full-text review.

For articles with an abstract or summary, we used a multistep process to screen for inclusion in the full-text review. First, 3 members of the study team (CTB, LB, and SM) independently screened a random sample of 6% (120/1897) of articles and discussed disagreements among the reviewers about whether articles should be included. We held a series of meetings to refine and finalize our screening criteria to improve agreement among our team. Second, we used single-reviewer screening to determine inclusion for the remaining 94% (1777/1897) of documents. Third, we used random dual review of a sample (445/1777, 25.04%) of documents that had only been reviewed by a single reviewer so that we could measure and report interrater agreement. Disagreements in inclusion decisions were resolved through consensus discussion by all 3 reviewers.

We decided to group issues and strategies using a 4-step framework that we adapted from previously published AI development pipeline literature sources [21,29-31]. The closest preexisting framework was described by Chen et al [29] as including 5 categories: Problem Selection, Data Collection, Outcome Definition, Algorithm Development, and Postdeployment Considerations. To make our results understandable to the broadest possible set of stakeholders, we expanded Chen et al's original "Problem Selection" category to include other aspects of the Background Context of AI development and use. We retained a category for issues related to Data Characteristics. We collapsed Outcome Definition together with Algorithm Development because they are related design decisions, and we renamed Postdeployment

Considerations to Deployment so that all forms of evaluation would be included. Thus, our 4 development categories in the framework became:

- **Background Context:** systemic and structural elements (eg, factors that influence Problem Selection). For Background Context, we defined systemic and structural elements as the societal and organizational characteristics influencing developers, including the rules and regulations in place at the local, regional, and national levels.
- **Data Characteristics:** quality and quantity of the data.
- **Design:** choice of model, variables, outcome definition, and objective function.
- **Deployment:** model evaluation, use, and maintenance.

Abstraction of Data Items for Issues and Strategies

Each article undergoing full-text review was reviewed by 1 of 3 members of the study team. Relevant citations listed in these articles were also reviewed to identify additional data sources. Our unit of analysis was an issue-strategy pair, defined as the linking of a particular equity issue to a potential strategy that could be used to improve equity for the AI application in health care. We defined an issue as a potential equity-related problem that had been suggested by at least one document author, and we defined a strategy as a recommended action to address an issue. We extracted issues and strategies named in each article using a data collection form consisting of the reference for each document, the specific issue(s) that the document discussed, and which strategies that the article proposed could be used to address the issue. Each document could include multiple issue-strategy pairs. We also abstracted the following items for each issue: narrative description of the issue, issue group (prespecified categories: Background Context, Data Characteristics, Design, and Deployment), representative quotes from the document, and representative quotes describing strategies. We included issues and strategies that were speculative or theoretical in addition to those that have been

“proven” to exist, because we believed this information would likely be valuable to developers and regulators who are interested in learning about emerging issues and solutions.

Synthesis of Results

We created our set of issues and strategies inductively: whenever an equity issue or strategy discussed in a document was not adequately described by the current set, we created a new entry. Definitions were refined in group meetings among the 3 members of the study team.

Ethics and Human Participants

The RAND Corporation Human Subjects Protection Committee (HSPC ID 2021-N0625), which functions as RAND’s

Institutional Review Board, determined that our study qualified for exemption from committee review.

Results From the Preliminary Stakeholder Consultation

Our stakeholders did not suggest any changes to the study topics proposed for our review. They suggested that we should include gray literature documents such as news articles, clinical trial protocols, and conference proceedings in our review in addition to peer-reviewed articles. Stakeholders also suggested that we investigate several topics related to AI and equity that they believed warranted further research ([Textbox 2](#)).

Textbox 2. Stakeholder recommendations for areas of focus in the scoping review.

Data sets, variable selection, and health equity

Stakeholders emphasized that there was a gap in current understanding about how limitations in training and validation data sets influenced AI application performance for vulnerable subpopulations and how strategies could be undertaken to protect such subpopulations. They also expressed concern that there was a tension in ensuring inclusion of underrepresented groups while also ensuring privacy for patients from such groups, and that strategies were needed to improve equity due to this tension.

Limitations in evaluating equity-related outcomes

Four interviewees suggested that it was important to investigate certain outcomes for vulnerable subgroups of patients, such as measures of cost, quality, and access to care, that might be challenging for developers to obtain.

Availability of equity-related information on AI algorithm performance

Four interviewees mentioned that AI may be used internally by an organization such as a health system or government agency, and that publicly available information about algorithm performance for vulnerable subgroups might be limited.

See [Multimedia Appendix 3](#) for additional results from the stakeholder consultation.

Results

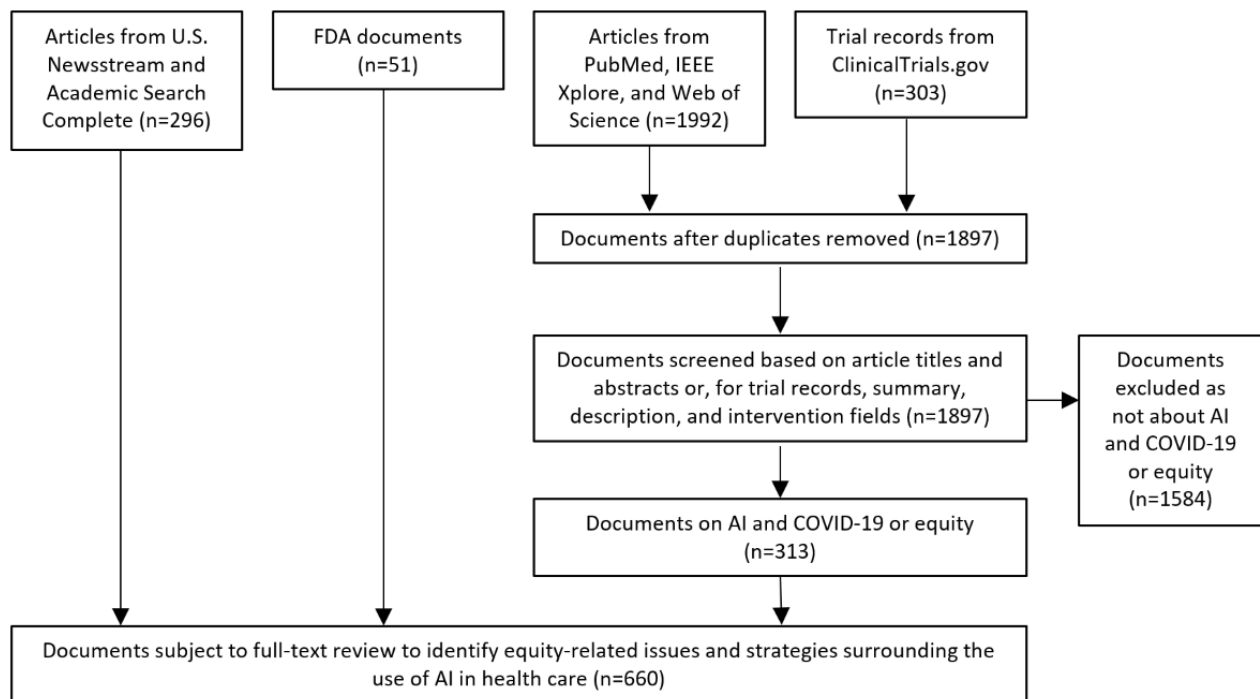
Search Output

Our search strategy identified a total of 2244 unique documents of potential interest. We conducted title and abstract review for 1897 documents or trial records, with 313 meeting inclusion criteria. For a 25% (445/1777) sample of records that were reviewed by 2 reviewers, interreviewer agreement on inclusion

was 88% (391/445; Cohen $\kappa=0.61$) [38]. We identified an additional 347 documents of interest that did not have abstracts to review, so they all underwent full-text review (296 news articles and 51 Food and Drug Administration documents).

In total, 660 documents meeting inclusion criteria underwent full-text review and were included in our analysis. The PRISMA flow diagram displaying the literature search and screening results is presented in [Figure 1](#) [36].

Figure 1. PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) flow diagram. AI: artificial intelligence; FDA: Food and Drug Administration.



Equity Issues and Strategies in Health AI

This section will present three tables and one figure that highlight the issues affecting equity for AI applications as well as the strategies we identified to address them.

We identified a total of 18 issues linked to 15 strategies. We present our main results in 2 parts. [Tables 1](#) and [2](#) display the

issues and strategies, respectively, that we identified in the literature, and we provide a brief narrative description for each item. Then, [Figure 2](#) and [Table 3](#) demonstrate how issues and strategies were linked together. The complete list of documents that identified each issue-strategy pair is provided in [Multimedia Appendix 4](#).

Table 1. Issues related to AI^a and health equity that were abstracted from the literature.

Category and issue	Description
Background Context	
Biased or nonrepresentative developers	Development team composition may be biased or poorly representative of the population, leading to mismatched priorities and blind spots.
Diminished accountability	Lack of developer accountability makes it difficult for individuals harmed by AI applications to obtain compensation.
Enabling discrimination	Developers may use AI algorithms to purposely discriminate for malice or for economic gain.
Data Characteristics	
Limited information on population characteristics	Insufficiently granular data on population characteristics may lead to inappropriately aggregating dissimilar groups, such as classifying race into only White and non-White.
Unrepresentative data or small sample sizes	Inadequate representation of groups in training data can lead to worse model performance in these groups, especially when training and deployment populations are poorly matched.
Bias ingrained in data	When data reflect past disparities or discrimination, algorithms may incorporate and perpetuate these patterns.
Inclusion of sensitive variables	Inclusion of sensitive information, such as race or income, may cause algorithms to inappropriately discriminate on these factors.
Exclusion of sensitive variables	Exclusion of sensitive information may reduce accuracy in some groups and lead to systematic bias due to a lack of explanatory power.
Limited reporting of information on protected groups	Lack of reporting on the composition of training data or model performance by group makes it difficult to know where to appropriately use models and whether they have disparate impacts.
Model Design	
Algorithms are not interpretable	When we do not understand why models make decisions, it is difficult to evaluate whether the decision-making approach is fair or equitable.
Optimizing algorithm accuracy and fairness may conflict	Optimizing models for fairness may introduce a trade-off between model accuracy and the fairness constraint, meaning that equity may come at the expense of decreased accuracy.
Ambiguity in and conflict among conceptions of equity	There are many conceptions of fairness and equity, which may be mutually exclusive or require sensitive data to evaluate.
Deployment Practices	
Proprietary algorithms or data unavailable for evaluation	When training data, model design, or the outputs of algorithms are proprietary, regulators and other independent evaluators may not be able to effectively assess risk of bias.
Overreliance on AI applications	Users may blindly trust algorithmic outputs, implementing decisions despite contrary evidence and perpetuating biases if the algorithm is discriminatory.
Underreliance on AI applications	People may be dismissive of algorithm outputs that challenge their own biases, thereby perpetuating discrimination.
Repurposing existing AI applications outside original scope	Models may be repurposed for use with new populations or to perform new functions without sufficient evaluation, bypassing safeguards on appropriate use.
Application development or implementation is rushed	Time constraints may exacerbate equity issues if they push developers to inappropriately repurpose existing models, use low-quality data, or skip validation.
Unequal access to AI	AI applications may be deployed more commonly in high-income areas, potentially amplifying preexisting disparities.

^aAI: artificial intelligence.

Table 2. Strategies to address AI^a equity issues that were abstracted from the literature.

Category and strategy	Description
Background Context	
Foster diversity	Create AI development teams with diverse characteristics, experiences, and roles to increase consideration of equity throughout development and decrease blind spots.
Train developers and users	Train AI developers and users in equity considerations and the ethical implications of AI, as these topics may be unfamiliar to some.
Engage the broader community	Foster community involvement throughout development, from conception to postdeployment, to increase the likelihood that developers prioritize equity concerns.
Improve governance	Enact robust regulation and industry standards to align AI applications with social norms, including equity, safety, and transparency.
Data Characteristics	
Improve diversity, quality, or quantity of data	Train models with large, diverse samples that are representative of the target population for the application and contain all relevant features.
Exclude sensitive variables to correct for bias	Exclude sensitive variables or replace them with variables that are more directly relevant to health outcomes to prevent models from discriminating directly on these characteristics.
Include sensitive variables to correct for bias	Include sensitive variables to improve model accuracy, increase explanatory power, and enable easier testing for inequitable impact.
Model Design	
Enforce fairness goals	Formulate a fairness norm and enforce it in the model by editing the input data, objective function, or model outputs.
Improve interpretability or explainability of the algorithm	Choose models that are inherently explainable (such as decision trees), build models with post hoc explainability, or explore explainable local approximations to model decision making.
Evaluate disparities in model performance	Evaluate model performance on a wide range of metrics across subgroups, particularly groups that might face inequitable impact, then report and act upon the results.
Use equity-focused checklists, guidelines, and similar tools	Incorporate equity-focused checklists into workflows for developers, reviewers of AI models, health care providers using an application, or patients who want to understand algorithm outputs.
Deployment Practices	
Increase model reporting and transparency	Provide more information on AI equity issues, including publishing standardized equity-related analyses on models, increasing independent model reviews, and requiring equity discussions in academic journals.
Seek or provide restitution for those negatively impacted by AI	Proactively provide restitution to those harmed by AI or create legal frameworks so they can seek restitution.
Avoid or reduce use of AI	Consider discontinuing model use if equity sequelae are severe or if improvement efforts have been fruitless.
Provide resources to those with less access to AI	Improve access to AI for disadvantaged groups and low-income countries by subsidizing infrastructure, creating education programs, or hosting AI conferences in these locations.

^aAI: artificial intelligence.

Figure 2. Issues related to AI and equity and strategies proposed to address them. The thickness and opacity of each line connecting an issue to a strategy are proportional to how frequently they were mentioned together. AI: artificial intelligence.

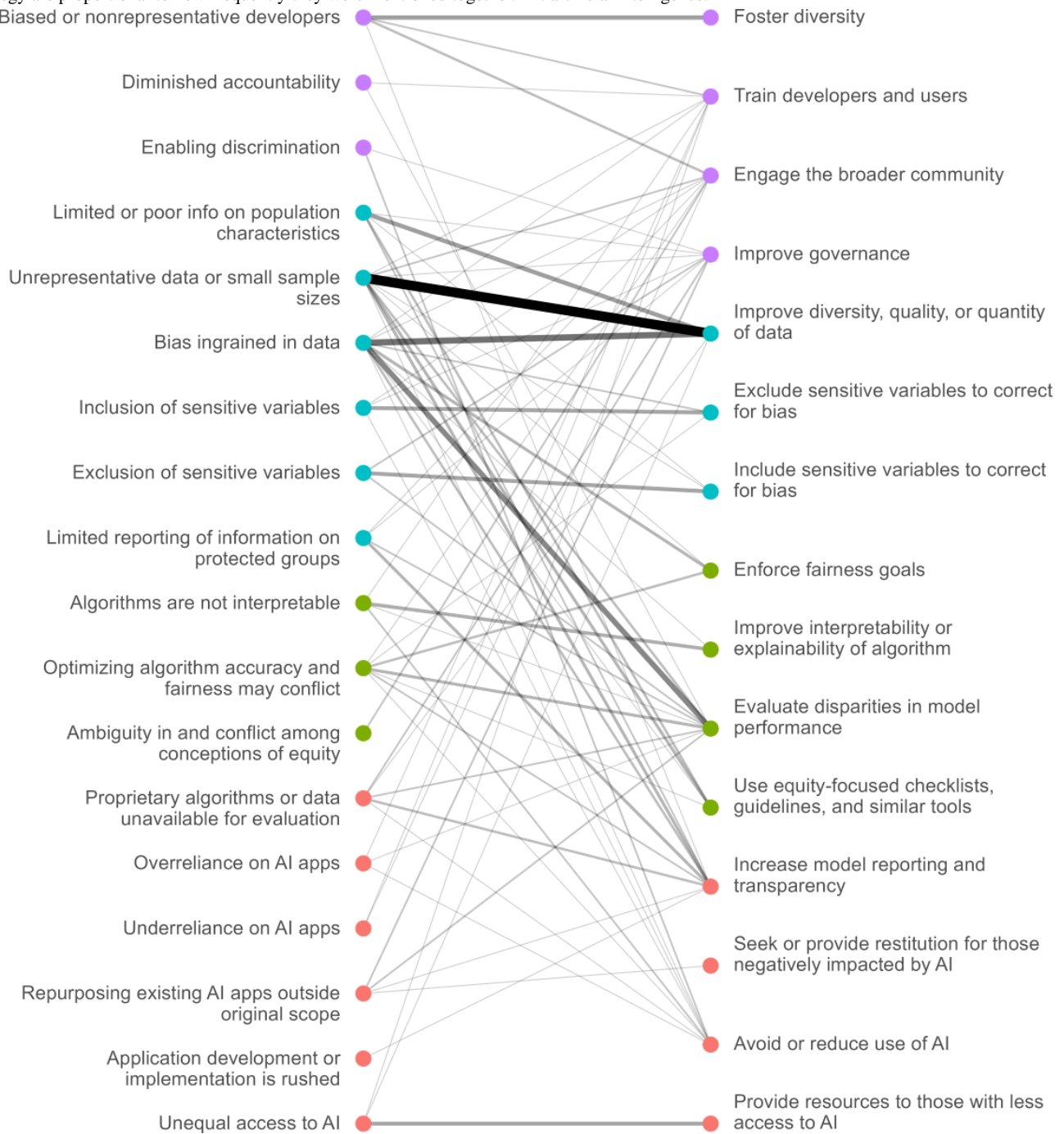


Table 3. The most common strategies mentioned in the literature for each health equity issue.

Category and issue	Issue frequency (N=195), n (%)	Most frequently linked strategy	Second most frequently linked strategy
Background Context			
Biased or nonrepresentative developers	13 (6.7)	Foster diversity	Engage the broader community
Diminished accountability	2 (1.0)	Evaluate disparities in model performance	Train developers and users
Enabling discrimination	3 (1.5)	Avoid or reduce use of AI ^a	Improve governance
Data Characteristics			
Limited information on population characteristics	14 (7.2)	Improve diversity, quality, or quantity of data	Use equity-focused checklists, guidelines, and similar tools
Unrepresentative data or small sample sizes	46 (23.6)	Improve diversity, quality, or quantity of data	Increase model reporting and transparency
Bias ingrained in data	37 (19.0)	Improve diversity, quality, or quantity of data	Evaluate disparities in model performance
Inclusion of sensitive variables	9 (4.6)	Exclude sensitive variables to correct for bias	Avoid or reduce use of AI
Exclusion of sensitive variables	10 (5.1)	Include sensitive variables to correct for bias	Evaluate disparities in model performance
Limited reporting of information on protected groups	8 (4.1)	Increase model reporting and transparency	Evaluate disparities in model performance
Model Design			
Algorithms are not interpretable	9 (4.6)	Improve interpretability or explainability of algorithm	Avoid or reduce use of AI
Optimizing algorithm accuracy and fairness may conflict	13 (6.7)	Evaluate disparities in model performance	Enforce fairness goals
Ambiguity in and conflict among conceptions of equity	2 (1.0)	Engage the broader community	— ^b
Deployment Practices			
Proprietary algorithms or data unavailable for evaluation	9 (4.6)	Increase model reporting and transparency	Evaluate disparities in model performance
Overreliance on AI applications	3 (1.5)	Avoid or reduce use of AI	Evaluate disparities in model performance
Underreliance on AI applications	2 (1.0)	Engage the broader community	Train developers and users
Repurposing existing AI applications outside original scope	6 (3.1)	Evaluate disparities in model performance	Improve governance
Application development or implementation is rushed	1 (0.5)	Increase model reporting and transparency	—
Unequal access to AI	8 (4.1)	Provide resources to those with less access to AI	Improve diversity, quality, or quantity of data

^aAI: artificial intelligence.

^bOnly 1 issue has been linked to the strategy.

Linking Issues and Strategies

In this section, we report how issues and strategies have been linked in the articles we reviewed. The strategies most frequently linked to each issue are shown in [Table 3](#), and the references provided in [Multimedia Appendix 2](#) offer more detail on how to apply a strategy to a given issue. A small number of issues comprise the majority of mentions in the literature: The top 5 issues constitute 63% (123/195, 63.1%) of all issue-strategy pairs. Each of these issues has several well-developed strategies,

usually focused on improving the quality of data or evaluating bias in model-decision making. By contrast, other issues are mentioned infrequently and do not have well-developed strategies. When only 1 issue has been linked to a strategy, the second column is presented with an em dash. We included an issue frequency column as a measure of how often issues have been mentioned in the literature.

[Figure 2](#) is a map of the 195 issue-strategy pairs identified in the literature, and it shows a complex many-to-many mapping

between issues and strategies in health equity and highlights which strategies and issues are most common. Each issue-strategy pair mentioned in the literature is shown as a link. Bolder lines indicate strategies and issues that are more frequently linked. A comprehensive list of links and the corresponding references is provided in [Multimedia Appendix 4](#). Out of the total 195 issue-strategy pairs, 50.3% (98/195) were identified in peer-reviewed literature. The remaining 49.7% (97/195) were from gray literature sources, including 14 conference proceedings, 11 news articles, 5 textbooks, 3 preprints, 2 press releases, 1 thesis, 1 clinical trial record, and 60 others (eg, reports and briefings).

Discussion

Principal Findings

By analyzing the literature on AI and health disparities we have identified 18 issues and 15 strategies that can be used to improve health equity in the realm of AI. Our work builds upon frameworks from the existing literature, identifying named strategies and issues for each stage of AI development and implementation. In addition, we draw 3 new insights from mapping the relationships between issues and strategies.

The framework published by Chen et al [29] offers 5 recommendations for improving equity, which can be paraphrased as follows: (1) problems should be tackled by diverse teams using frameworks that increase the probability that equity will be achieved; (2) data collection should be framed as an important front-of-mind concern, including encouragement of disclosing imbalanced data sets; (3) outcome choice should reflect the task at hand in an unbiased manner; (4) developers should reflect on the goals of the model during development and preanalysis; and (5) audits should be designed to identify specific harms, including harms at the level of the group rather than at the population. While these are important and sound recommendations, our results additionally emphasize the need to engage with communities throughout the development and deployment phases, identify opportunities for equity-focused governance at the local and national levels, and identify additional opportunities for improvement after algorithms are found to impair equity (eg, avoiding or reducing AI use, providing resources to those with less access to AI, and providing restitution to those negatively impacted by AI). Our comprehensive mapping of issues and strategies can be useful to stakeholders of all types, including developers, representatives of vulnerable groups, and regulators.

The Literature Focuses on a Small Set of Issues

A small set of issues dominates the literature. The top 5 issues comprise nearly two thirds of all issue-strategy pairs. The discourse around health AI equity focuses on Data Characteristics: almost two-thirds of all issue-strategy pairs are related to data. These issues are widely researched, and, therefore, we encountered many corresponding strategies to address them. Some strategies directly address data quality, while others accept data limitations and try to improve fairness despite poor data quality.

Much of the discourse on model design focuses on the trade-off between accuracy and fairness [39-43]. This multifaceted problem requires that stakeholders select a definition of fairness and analyze how accuracy/fairness trade-offs will balance in specific applications. The most common approach to improving model design involves measuring disparities in model performance and revising the model to enforce fairness goals [44]. As definitions of fairness may conflict, developers and evaluators should test the impact of different constraints across a broad range of metrics (such as accuracy, false-positive rate, and false-negative rate) and report group-level disparities in each of these metrics [45]. Equity-relevant model design literature is most developed for classification or regression tasks, and there is less guidance in other areas such as online learning [46]. Relevant subgroups are often application specific, and the data on these subgroups may not be available [47].

Other issues were rarely discussed and have a limited number of associated strategies. For example, several issues reflect concerns about how AI is deployed—especially when AI applications are used outside their original scope or when they are rushed through development and into production without sufficient testing.

Even if an issue is not frequently discussed in the literature, it may still be important. In other words, an issue may not be discussed frequently because there is limited evidence of equity impact or because corresponding strategies are underdeveloped. We believe that some issues may have been insufficiently discussed despite their promise as topics that would benefit from future research. For example, future work is warranted to investigate the negative impacts of the following issues: repurposing AI applications outside their original scope, inadequate descriptions of population characteristics, and lack of accountability for the unintended consequences of AI on health equity.

Strategies Are Multipurpose

While some strategies, such as improving interpretability, are tailored to specific issues, most strategies are multipurpose. The top 5 most frequently mentioned strategies, which account for more than half of issue-strategy pairs in our sample, are collectively linked to all 18 issues. Each of these strategies is linked to critical aspects of application development. Evaluating disparities in model performance is often necessary for quantifying bias across subgroups. Similarly, improving data is important across a broad range of issues because the decision-making logic of AI models flows directly from training data. Community engagement and improved governance can increase the consideration of equity issues throughout all stages of AI algorithm development. Community stakeholders should be involved at all stages of production, including deciding whether an application should be built, setting goals for the model, defining fairness [48], and guarding against unintended consequences after deployment [21,49-51]. Improving governance is usually advocated in the form of guiding principles for AI use [25,52] or “soft governance” such as industry-organized protocols [53,54]. Regulation is not frequently advocated, although it is unclear whether this is

because researchers believe regulation would be ineffective or because they prefer to focus on technical solutions.

Small Sets of Strategies Can Address a Broad Set of Issues

Sometimes it is only practical to focus on a small set of strategies. For instance, in their Algorithmic Bias Playbook, Obermeyer et al [32] suggested that organizations identify biased algorithms and then retrain them on less biased targets, improve the representativeness of their data set, or consider discontinuing their use.

Once stakeholders have identified issues that are relevant for a specific application, they can use Table 3 and Figure 2 to select a set of strategies to address them. The most common 5 strategies cited above are a good starting point because of their broad coverage of issues. However, not all these strategies may be feasible, and others may require complementation with additional strategies to fully address a specific issue.

Consider an example use case for our mapping of equity-relevant issues and strategies to address them: A developer has been commissioned to build an open-source predictive model of emergency department admission probability based on electronic health records. The developer has identified data issues related to bias and representativeness, but is also concerned that the model may be less accurate for some subgroups of patients. The developer may consider the top 5 most common strategies first, and then may realize that modifying the data collection process is infeasible. Although improving governance does not necessarily require new legislative or regulatory action, it does involve collective action between industry and the broader community, so it may seem feasible in certain scenarios. However, the remaining 3 of the top 5 strategies can be implemented by a single stakeholder without coordinating collective action across different groups. Anyone with model access and demographic data can evaluate disparities in model performance and increase model reporting and transparency. Similarly, all developers can seek input from affected communities when they begin the development process.

The developer could then use Figure 2 to select a set of complementary strategies specific to some of the issues. If their evaluation did find disparate performance across groups, then they could enforce fairness constraints in the input data, model design, or model outputs. They may also review the model using an equity-focused checklist, such as the Prediction Model Risk of Bias Assessment Tool (PROBAST) [55], as this is low-cost and may identify other avenues to improve equity. They may also decide that they can better engage with the relevant stakeholders if they can explain the model's decision-making processes and develop model report cards for equity.

After completing this exercise, the developer will have identified an initial set of strategies that is within their scope of action. This set may evolve over time, especially as the broader community is engaged: For example, community stakeholders may help identify important features the developer overlooked (such as social determinants of health), suggest different definitions of equity, or question whether AI should be used at all [56].

This use-case example is one approach to addressing a complex set of equity issues. For most AI applications, we expect that developers will be able to identify a small set of strategies to address a broad range of equity issues. Particularly important issues may require multiple complementary strategies. We recommend that developers start by considering which of the 5 most common strategies are suitable for an application and then adding additional complementary strategies as needed—particularly low-cost strategies such as the use of the PROBAST checklist.

Limitations

This scoping review has several limitations. First, due to space constraints, the descriptions of each issue and strategy are brief. This means that stakeholders may need to access additional resources to take action and operationalize a strategy. For instance, if enforcing fairness goals is identified as a useful strategy, stakeholders need to decide what fairness rule to use and how to modify data inputs, the model objective function, or model outputs [21,57-60]. To better understand issues and strategies, stakeholders should use Multimedia Appendix 4 to find relevant documents. More detailed descriptions of issues and strategies will also be available in a subsequent report that will be published by the funder of this study, the Patient-Centered Outcomes Research Institute.

Second, some issues and strategies may conflict. For example, both inclusion and exclusion of sensitive variables are discussed as having either a positive or a negative influence on the impact of health AI on equity, depending on context and perspective. As a result, we include these as both issues and strategies in our study, reflecting the unsettled and context-dependent nature of debate on this topic within the literature.

Third, our search strategy included gray literature sources, so some of the issue-strategy pairs are likely to be speculative rather than proven to be effective. Out of 195 issue-strategy pairings, 98 were from peer-reviewed literature and 97 were from gray literature sources such as reports, news articles, conference proceedings, and preprint articles. Readers should consult the sources of the issue-strategy pairs when determining whether a given strategy should be used.

Fourth, we did not rate the quality of issues, strategies, or the articles from which we identified issue-strategy pairs. Some sources go into detail about health equity issues and strategies, others only make general recommendations or may represent outmoded views. The goal of this scoping review was to identify which issues and strategies are highlighted in the literature. Future reviews could instead focus on identifying the best or most developed strategies.

Fifth, the issues and strategies we identified are not entirely distinct: some are intermediaries that lead to other issues or strategies. For instance, repurposing an application is not inherently inequitable, but may increase the chance that the training data are unrepresentative of the target population. Similarly, uninterpretable algorithms do not create biased outcomes, but make them more difficult to detect. The same applies to strategies: using equity checklists does not directly solve problems, but makes it more likely that developers identify

equity issues and appropriate strategies. We included these intermediary issues and strategies because they provide a richer description of intervention points for promoting health equity.

Sixth, there are other prominent concerns about AI and equity that were out of scope for our review. For example, AI applications may displace human workers in ways that could increase economic and health disparities, or the default use of female voices in AI assistants that perform clerical tasks may perpetuate bias and lead to negative effects on health equity for women [51]. While these concerns are raised in the context of economic or social disparities, we found no discussion of their impact on health equity specifically, and thus did not include them in our study.

Conclusions

Our work contributes to a growing body of AI health equity literature. We add to this literature by creating a many-to-many mapping between strategies and issues and by reviewing the literature to identify how often each strategy is linked to each issue. This scoping review is useful for a wide array of stakeholders, including developers, users, policymakers, and researchers who may wish to implement strategies to improve health equity for vulnerable populations of interest. While no set of strategies can eliminate the equity concerns posed by health AI, small sets of strategies can often mitigate many of the most pressing issues. We should also recognize that existing nonalgorithmic decision making is imperfect. By thoughtfully adopting complementary sets of strategies that cover a broad range of equity issues, AI models may offer improvements in equity over the status quo.

Acknowledgments

This work was funded by the Patient-Centered Outcomes Research Institute (PCORI) under Contract No. IDIQTO#22-RAND-ENG-AOSEPP-04-01-2020. All statements, findings, and conclusions in this publication are solely those of the authors and do not necessarily represent the views of the Patient-Centered Outcomes Research Institute (PCORI). The funders advised on the study design; the funders played no role in data collection and analysis, decision to publish, or preparation of the manuscript.

Conflicts of Interest

None declared.

Multimedia Appendix 1

PRISMA-ScR (Preferred Reporting Items for Systematic Reviews and Meta-Analyses Extension for Scoping Reviews) checklist. [[DOCX File, 86 KB - ai_v2i1e42936_app1.docx](#)]

Multimedia Appendix 2

Literature search documentation and stakeholder interview protocol. [[DOCX File, 90 KB - ai_v2i1e42936_app2.docx](#)]

Multimedia Appendix 3

Results from stakeholder consultation. [[DOCX File, 25 KB - ai_v2i1e42936_app3.docx](#)]

Multimedia Appendix 4

Table of documents linking AI health equity issues and strategies. [[DOCX File, 70 KB - ai_v2i1e42936_app4.docx](#)]

References

1. Secinaro S, Calandra D, Secinaro A, Muthurangu V, Biancone P. The role of artificial intelligence in healthcare: a structured literature review. *BMC Med Inform Decis Mak* 2021 Apr 10;21(1):125 [[FREE Full text](#)] [doi: [10.1186/s12911-021-01488-9](https://doi.org/10.1186/s12911-021-01488-9)] [Medline: [33836752](https://pubmed.ncbi.nlm.nih.gov/33836752/)]
2. Yin J, Ngiam KY, Teo HH. Role of Artificial Intelligence Applications in Real-Life Clinical Practice: Systematic Review. *J Med Internet Res* 2021 Apr 22;23(4):e25759 [[FREE Full text](#)] [doi: [10.2196/25759](https://doi.org/10.2196/25759)] [Medline: [33885365](https://pubmed.ncbi.nlm.nih.gov/33885365/)]
3. Noorbakhsh-Sabet N, Zand R, Zhang Y, Abedi V. Artificial Intelligence Transforms the Future of Health Care. *Am J Med* 2019 Jul;132(7):795-801 [[FREE Full text](#)] [doi: [10.1016/j.amjmed.2019.01.017](https://doi.org/10.1016/j.amjmed.2019.01.017)] [Medline: [30710543](https://pubmed.ncbi.nlm.nih.gov/30710543/)]
4. Rajpurkar P, Chen E, Banerjee O, Topol EJ. AI in health and medicine. *Nat Med* 2022 Jan;28(1):31-38. [doi: [10.1038/s41591-021-01614-0](https://doi.org/10.1038/s41591-021-01614-0)] [Medline: [35058619](https://pubmed.ncbi.nlm.nih.gov/35058619/)]
5. Topol E. *Deep medicine: how artificial intelligence can make healthcare human again*. New York, NY: Basic Books; 2019.

6. Rajkomar A, Dean J, Kohane I. Machine Learning in Medicine. *N Engl J Med* 2019 Apr 04;380(14):1347-1358. [doi: [10.1056/NEJMra1814259](https://doi.org/10.1056/NEJMra1814259)] [Medline: [30943338](https://pubmed.ncbi.nlm.nih.gov/30943338/)]
7. Gunasekeran DV, Tseng RMWW, Tham Y, Wong TY. Applications of digital health for public health responses to COVID-19: a systematic scoping review of artificial intelligence, telehealth and related technologies. *NPJ Digit Med* 2021 Feb 26;4(1):40 [FREE Full text] [doi: [10.1038/s41746-021-00412-9](https://doi.org/10.1038/s41746-021-00412-9)] [Medline: [33637833](https://pubmed.ncbi.nlm.nih.gov/33637833/)]
8. Wynants L, Van Calster B, Collins GS, Riley RD, Heinze G, Schuit E, et al. Prediction models for diagnosis and prognosis of covid-19: systematic review and critical appraisal. *BMJ* 2020 Apr 07;369:m1328 [FREE Full text] [doi: [10.1136/bmj.m1328](https://doi.org/10.1136/bmj.m1328)] [Medline: [32265220](https://pubmed.ncbi.nlm.nih.gov/32265220/)]
9. Musulin J, Baressi Šegota S, Štifanić D, Lorencin I, Anđelić N, Šušteršič T, et al. Application of Artificial Intelligence-Based Regression Methods in the Problem of COVID-19 Spread Prediction: A Systematic Review. *Int J Environ Res Public Health* 2021 Apr 18;18(8):4287 [FREE Full text] [doi: [10.3390/ijerph18084287](https://doi.org/10.3390/ijerph18084287)] [Medline: [33919496](https://pubmed.ncbi.nlm.nih.gov/33919496/)]
10. Rasheed J, Jamil A, Hameed AA, Al-Turjman F, Rasheed A. COVID-19 in the Age of Artificial Intelligence: A Comprehensive Review. *Interdiscip Sci* 2021 Jun;13(2):153-175 [FREE Full text] [doi: [10.1007/s12539-021-00431-w](https://doi.org/10.1007/s12539-021-00431-w)] [Medline: [33886097](https://pubmed.ncbi.nlm.nih.gov/33886097/)]
11. Jamshidi M, Roshani S, Talla J, Lalbakhsh A, Peroutka Z, Roshani S, et al. A Review of the Potential of Artificial Intelligence Approaches to Forecasting COVID-19 Spreading. *AI* 2022 May 19;3(2):493-511 [FREE Full text] [doi: [10.3390/ai3020028](https://doi.org/10.3390/ai3020028)]
12. Malik YS, Sircar S, Bhat S, Ansari MI, Pande T, Kumar P, et al. How artificial intelligence may help the Covid-19 pandemic: Pitfalls and lessons for the future. *Rev Med Virol* 2021 Sep;31(5):1-11 [FREE Full text] [doi: [10.1002/rmv.2205](https://doi.org/10.1002/rmv.2205)] [Medline: [33476063](https://pubmed.ncbi.nlm.nih.gov/33476063/)]
13. Rasheed J, Jamil A, Hameed AA, Aftab U, Aftab J, Shah SA, et al. A survey on artificial intelligence approaches in supporting frontline workers and decision makers for the COVID-19 pandemic. *Chaos Solitons Fractals* 2020 Dec;141:110337 [FREE Full text] [doi: [10.1016/j.chaos.2020.110337](https://doi.org/10.1016/j.chaos.2020.110337)] [Medline: [33071481](https://pubmed.ncbi.nlm.nih.gov/33071481/)]
14. Rasheed J, Shubair RM. Screening Lung Diseases Using Cascaded Feature Generation and Selection Strategies. *Healthcare (Basel)* 2022 Jul 14;10(7):1313 [FREE Full text] [doi: [10.3390/healthcare10071313](https://doi.org/10.3390/healthcare10071313)] [Medline: [35885839](https://pubmed.ncbi.nlm.nih.gov/35885839/)]
15. Rasheed J. Analyzing the Effect of Filtering and Feature-Extraction Techniques in a Machine Learning Model for Identification of Infectious Disease Using Radiography Imaging. *Symmetry* 2022 Jul 07;14(7):1398. [doi: [10.3390/sym14071398](https://doi.org/10.3390/sym14071398)]
16. Pérez-Stable EJ, Jean-Francois B, Aklin CF. Leveraging Advances in Technology to Promote Health Equity. *Med Care* 2019 Jun;57 Suppl 6 Suppl 2:S101-S103. [doi: [10.1097/MLR.0000000000001112](https://doi.org/10.1097/MLR.0000000000001112)] [Medline: [31095045](https://pubmed.ncbi.nlm.nih.gov/31095045/)]
17. Veinot TC, Ancker JS, Bakken S. Health informatics and health equity: improving our reach and impact. *J Am Med Inform Assoc* 2019 Aug 01;26(8-9):689-695 [FREE Full text] [doi: [10.1093/jamia/ocz132](https://doi.org/10.1093/jamia/ocz132)] [Medline: [31411692](https://pubmed.ncbi.nlm.nih.gov/31411692/)]
18. Chen IY, Szolovits P, Ghassemi M. Can AI Help Reduce Disparities in General Medical and Mental Health Care? *AMA J Ethics* 2019 Feb 01;21(2):E167-E179 [FREE Full text] [doi: [10.1001/amajethics.2019.167](https://doi.org/10.1001/amajethics.2019.167)] [Medline: [30794127](https://pubmed.ncbi.nlm.nih.gov/30794127/)]
19. Chen IY, Joshi S, Ghassemi M. Treating health disparities with artificial intelligence. *Nat Med* 2020 Jan 13;26(1):16-17. [doi: [10.1038/s41591-019-0649-2](https://doi.org/10.1038/s41591-019-0649-2)] [Medline: [31932779](https://pubmed.ncbi.nlm.nih.gov/31932779/)]
20. Thomasian NM, Eickhoff C, Adashi EY. Advancing health equity with artificial intelligence. *J Public Health Policy* 2021 Dec 22;42(4):602-611 [FREE Full text] [doi: [10.1057/s41271-021-00319-5](https://doi.org/10.1057/s41271-021-00319-5)] [Medline: [34811466](https://pubmed.ncbi.nlm.nih.gov/34811466/)]
21. Rajkomar A, Hardt M, Howell MD, Corrado G, Chin MH. Ensuring Fairness in Machine Learning to Advance Health Equity. *Ann Intern Med* 2018 Dec 18;169(12):866-872 [FREE Full text] [doi: [10.7326/M18-1990](https://doi.org/10.7326/M18-1990)] [Medline: [30508424](https://pubmed.ncbi.nlm.nih.gov/30508424/)]
22. Panch T, Mattie H, Celi LA. The "inconvenient truth" about AI in healthcare. *NPJ Digit Med* 2019;2:77 [FREE Full text] [doi: [10.1038/s41746-019-0155-4](https://doi.org/10.1038/s41746-019-0155-4)] [Medline: [31453372](https://pubmed.ncbi.nlm.nih.gov/31453372/)]
23. Panch T, Pearson-Stuttard J, Greaves F, Atun R. Artificial intelligence: opportunities and risks for public health. *Lancet Digit Health* 2019 May;1(1):e13-e14 [FREE Full text] [doi: [10.1016/S2589-7500\(19\)30002-0](https://doi.org/10.1016/S2589-7500(19)30002-0)] [Medline: [33323236](https://pubmed.ncbi.nlm.nih.gov/33323236/)]
24. Parikh RB, Teeple S, Navathe AS. Addressing Bias in Artificial Intelligence in Health Care. *JAMA* 2019 Dec 24;322(24):2377-2378. [doi: [10.1001/jama.2019.18058](https://doi.org/10.1001/jama.2019.18058)] [Medline: [31755905](https://pubmed.ncbi.nlm.nih.gov/31755905/)]
25. World Health Organization. Ethics and Governance of Artificial Intelligence for Health: WHO Guidance. Geneva, Switzerland: World Health Organization; 2021.
26. Mhasawade V, Zhao Y, Chunara R. Machine learning and algorithmic fairness in public and population health. *Nat Mach Intell* 2021 Jul 29;3(8):659-666. [doi: [10.1038/s42256-021-00373-4](https://doi.org/10.1038/s42256-021-00373-4)]
27. World Health Organization. Social determinants of health: Health equity. Geneva, Switzerland: World Health Organization; 2022 Nov 13. URL: <https://www.who.int/health-topics/social-determinants-of-health> [accessed 2023-01-09]
28. Braveman P, Arkin E, Orleans T, Proctor D, Plough A. What is health equity? And what difference does a definition make? National Collaborating Centre for Determinants of Health. Princeton, NJ: Robert Wood Johnson Foundation; 2017. URL: <https://nccdh.ca/resources/entry/what-is-health-equity-and-what-difference-does-a-definition-make> [accessed 2023-01-09]
29. Chen IY, Pierson E, Rose S, Joshi S, Ferryman K, Ghassemi M. Ethical Machine Learning in Healthcare. *Annu Rev Biomed Data Sci* 2021 Jul;4:123-144. [doi: [10.1146/annurev-biodatasci-092820-114757](https://doi.org/10.1146/annurev-biodatasci-092820-114757)] [Medline: [34396058](https://pubmed.ncbi.nlm.nih.gov/34396058/)]
30. Vokinger KN, Feuerriegel S, Kesselheim AS. Mitigating bias in machine learning for medicine. *Commun Med (Lond)* 2021 Aug 23;1:25. [doi: [10.1038/s43856-021-00028-w](https://doi.org/10.1038/s43856-021-00028-w)] [Medline: [34522916](https://pubmed.ncbi.nlm.nih.gov/34522916/)]

31. Yeung D, Khan I, Kalra N, Osoba O. Identifying Systemic Bias in the Acquisition of Machine Learning Decision Aids for Law Enforcement Applications. RAND. Santa Monica, CA: RAND Corporation; 2021. URL: <https://www.rand.org/pubs/perspectives/PEA862-1.html> [accessed 2023-01-09]
32. Obermeyer Z, Nissan R, Stern M, Eaneff S, Bembeneck E, Mullainathan S. Algorithmic Bias Playbook. Federal Trade Commission. Chicago (Chicago Booth), Illinois: The Center for Applied Artificial Intelligence; 2021. URL: https://www.ftc.gov/system/files/documents/public_events/1582978/algorithmic-bias-playbook.pdf [accessed 2023-01-09]
33. O'Brien N, Van Dael J, Clarke K, Gardner C, O'Shaughnessy J, Darzi A, et al. Addressing racial and ethnic inequities in data-driven health technologies. Imperial College London. 2022 Feb 24. URL: https://spiral.imperial.ac.uk/bitstream/10044/1/94902/2/Imperial_IGHI_AddressRacialandEthnicInequities%20Report.pdf [accessed 2022-12-14]
34. Mann S, Berdahl CT, Baker L, Girosi F. Artificial intelligence applications used in the clinical response to COVID-19: A scoping review. PLOS Digit Health 2022 Oct 17;1(10):e0000132. [doi: [10.1371/journal.pdig.0000132](https://doi.org/10.1371/journal.pdig.0000132)]
35. Arksey H, O'Malley L. Scoping studies: towards a methodological framework. International Journal of Social Research Methodology 2005 Feb;8(1):19-32. [doi: [10.1080/1364557032000119616](https://doi.org/10.1080/1364557032000119616)]
36. Tricco AC, Lillie E, Zarin W, O'Brien KK, Colquhoun H, Levac D, et al. PRISMA Extension for Scoping Reviews (PRISMA-ScR): Checklist and Explanation. Ann Intern Med 2018 Oct 02;169(7):467-473 [FREE Full text] [doi: [10.7326/M18-0850](https://doi.org/10.7326/M18-0850)] [Medline: [30178033](https://pubmed.ncbi.nlm.nih.gov/30178033/)]
37. Booth A, Carroll C. Systematic searching for theory to inform systematic reviews: is it feasible? Is it desirable? Health Info Libr J 2015 Sep;32(3):220-235 [FREE Full text] [doi: [10.1111/hir.12108](https://doi.org/10.1111/hir.12108)] [Medline: [26095232](https://pubmed.ncbi.nlm.nih.gov/26095232/)]
38. McHugh ML. Interrater reliability: the kappa statistic. Biochem Med (Zagreb) 2012;22(3):276-282 [FREE Full text] [Medline: [23092060](https://pubmed.ncbi.nlm.nih.gov/23092060/)]
39. Rodolfa KT, Lamba H, Ghani R. Empirical observation of negligible fairness-accuracy trade-offs in machine learning for public policy. Nat Mach Intell 2021 Oct 14;3(10):896-904. [doi: [10.1038/s42256-021-00396-x](https://doi.org/10.1038/s42256-021-00396-x)]
40. Desiere S, Struyven L. Using Artificial Intelligence to classify Jobseekers: The Accuracy-Equity Trade-off. J. Soc. Pol 2020 May 08;50(2):367-385. [doi: [10.1017/s0047279420000203](https://doi.org/10.1017/s0047279420000203)]
41. Dutta S, Wei D, Yueksel H, Chen P, Liu S, Varshney K. Is there a trade-off between fairness and accuracy? a perspective using mismatched hypothesis testing. In: Proceedings of the 37th International Conference on Machine Learning (PMLR), Vol. 119. 2020 Presented at: Proceedings of Machine Learning Research; July 13-18, 2020; Online p. 2803-2813 URL: <http://proceedings.mlr.press/v119/dutta20a.html>
42. Cooper A, Abrams E, Na N, editors. Emergent unfairness in algorithmic fairness-accuracy trade-off research. In: AIES '21: Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society. New York, NY: Association for Computing Machinery; 2021 Presented at: AIES '21: AAAI/ACM Conference on AI, Ethics, and Society; May 19-21, 2021; Virtual p. 46-54. [doi: [10.1145/3461702.3462519](https://doi.org/10.1145/3461702.3462519)]
43. Liu S, Vicente LN. Accuracy and fairness trade-offs in machine learning: a stochastic multi-objective approach. Comput Manag Sci 2022 Apr 21;19(3):513-537. [doi: [10.1007/s10287-022-00425-z](https://doi.org/10.1007/s10287-022-00425-z)]
44. Paul A, Jolley C, Anthony A. Reflecting the Past, Shaping the Future: Making AI Work for International Development. USAID. 2018. URL: <https://www.usaid.gov/digital-development/machine-learning/AI-ML-in-development> [accessed 2023-01-10]
45. Rodolfa K, Saleiro P, Ghani R. Bias and Fairness (Chapter 11). In: Big Data and Social Science. London, UK: Chapman & Hall/CRC; 2020:1-32.
46. Chouldechova A, Roth A. The frontiers of fairness in machine learning. arXiv. Preprint posted online on October 20, 2018 [FREE Full text]
47. Holstein K, Wortman VJ, Daumé IH, Dudik M, Wallach H. Improving Fairness in Machine Learning Systems: What Do Industry Practitioners Need? In: CHI '19: Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems. 2019 Presented at: CHI '19: CHI Conference on Human Factors in Computing Systems; May 4-9, 2019; Glasgow, Scotland, UK p. 1-16. [doi: [10.1145/3290605.3300830](https://doi.org/10.1145/3290605.3300830)]
48. Horvitz E, Clyburn M, Griffiths JM, Matheny J. Privacy and Ethics Recommendations for Computing Applications Developed to Mitigate COVID-19: White Paper Series on Pandemic Response and Preparedness. UNT Digital Library. Washington, DC: National Security Commission on Artificial Intelligence (U.S.); 2020. URL: <https://digital.library.unt.edu/ark:/67531/metadc1851194/> [accessed 2023-01-09]
49. Zimmer M, Franco Z, Madiraju P, Echeveste C, Heindel K, Ogle J. Public Opinion Research on Artificial Intelligence in Public Health Responses: Results of Focus Groups with Four Communities. AAAS. Washington, DC: AAAS Center for Public Engagement with Science and Technology; 2021 Aug 10. URL: <https://www.aaas.org/sites/default/files/2021-09/AI%20in%20Public%20Health%20Focus%20Groups%20-%20Final%20Report%20with%20Appendix.pdf> [accessed 2023-01-09]
50. Morley J, Machado CCV, Burr C, Cowls J, Joshi I, Taddeo M, et al. The ethics of AI in health care: A mapping review. Soc Sci Med 2020 Sep;260:113172. [doi: [10.1016/j.socscimed.2020.113172](https://doi.org/10.1016/j.socscimed.2020.113172)] [Medline: [32702587](https://pubmed.ncbi.nlm.nih.gov/32702587/)]
51. Shachar C, Gerke S, Adashi EY. AI Surveillance during Pandemics: Ethical Implementation Imperatives. Hastings Cent Rep 2020 May;50(3):18-21 [FREE Full text] [doi: [10.1002/hast.1125](https://doi.org/10.1002/hast.1125)] [Medline: [32596887](https://pubmed.ncbi.nlm.nih.gov/32596887/)]
52. Google. Our Principles ? Google AI. Google. URL: <https://ai.google/principles/> [accessed 2022-01-09]

53. Osoba OA, Boudreaux B, Saunders J, Irwin JL, Mueller PA, Cherney S. Algorithmic Equity: A Framework for Social Applications. Santa Monica, CA: RAND Corporation; 2019.
54. Villasenor J. Soft law as a complement to AI regulation. Brookings Institution. 2020 Jul 31. URL: <https://www.brookings.edu/research/soft-law-as-a-complement-to-ai-regulation/> [accessed 2022-12-29]
55. Wolff RF, Moons KGM, Riley RD, Whiting PF, Westwood M, Collins GS, PROBAST Group†. PROBAST: A Tool to Assess the Risk of Bias and Applicability of Prediction Model Studies. *Ann Intern Med* 2019 Jan 01;170(1):51-58 [FREE Full text] [doi: [10.7326/M18-1376](https://doi.org/10.7326/M18-1376)] [Medline: [30596875](https://pubmed.ncbi.nlm.nih.gov/30596875/)]
56. Miller K. When Algorithmic Fairness Fixes Fail: The Case for Keeping Humans in the Loop: Stanford University Human-Centered Artificial Intelligence; November 2, 2020. Stanford. 2020 Nov 2. URL: <https://hai.stanford.edu/news/when-algorithmic-fairness-fixes-fail-case-keeping-humans-loop> [accessed 2022-03-08]
57. Bellamy RKE, Dey K, Hind M, Hoffman SC, Houde S, Kannan K, et al. AI Fairness 360: An extensible toolkit for detecting and mitigating algorithmic bias. *IBM J. Res. & Dev* 2019 Jul 1;63(4/5):4:1-4:15. [doi: [10.1147/jrd.2019.2942287](https://doi.org/10.1147/jrd.2019.2942287)]
58. Wawira Gichoya J, McCoy LG, Celi LA, Ghassemi M. Equity in essence: a call for operationalising fairness in machine learning for healthcare. *BMJ Health Care Inform* 2021 Apr;28(1):e100289 [FREE Full text] [doi: [10.1136/bmjhci-2020-100289](https://doi.org/10.1136/bmjhci-2020-100289)] [Medline: [33910923](https://pubmed.ncbi.nlm.nih.gov/33910923/)]
59. Parbhoo S, Wawira Gichoya J, Celi LA, de la Hoz M, for MIT Critical Data. Operationalising fairness in medical algorithms. *BMJ Health Care Inform* 2022 Jun;29(1):e100617 [FREE Full text] [doi: [10.1136/bmjhci-2022-100617](https://doi.org/10.1136/bmjhci-2022-100617)] [Medline: [35688512](https://pubmed.ncbi.nlm.nih.gov/35688512/)]
60. Chin C, Robison M. How AI bots and voice assistants reinforce gender bias. Brookings Institution. 2020. URL: <https://www.brookings.edu/research/how-ai-bots-and-voice-assistants-reinforce-gender-bias/> [accessed 2022-11-29]

Abbreviations

AI: artificial intelligence

BeHEMoTh: Behavior, Health condition, Exclusions, Models, or Theories

FDA: Food and Drug Administration

IEEE: Institute of Electrical and Electronics Engineers

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

PROBAST: Prediction Model Risk of Bias Assessment Tool

Edited by G Eysenbach, K El Emam, B Malin; submitted 24.09.22; peer-reviewed by I Iyamu, J Ancker, J Rasheed, S Feuerriegel; comments to author 19.10.22; revised version received 14.12.22; accepted 29.12.22; published 07.02.23.

Please cite as:

Berdahl CT, Baker L, Mann S, Osoba O, Giroso F

Strategies to Improve the Impact of Artificial Intelligence on Health Equity: Scoping Review

JMIR AI 2023;2:e42936

URL: <https://ai.jmir.org/2023/1/e42936>

doi: [10.2196/42936](https://doi.org/10.2196/42936)

PMID:

©Carl Thomas Berdahl, Lawrence Baker, Sean Mann, Osonde Osoba, Federico Giroso. Originally published in JMIR AI (<https://ai.jmir.org>), 07.02.2023. This is an open-access article distributed under the terms of the Creative Commons Attribution License (<https://creativecommons.org/licenses/by/4.0/>), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR AI, is properly cited. The complete bibliographic information, a link to the original publication on <https://www.ai.jmir.org/>, as well as this copyright and license information must be included.

Original Paper

Forecasting Artificial Intelligence Trends in Health Care: Systematic International Patent Analysis

Stan Benjamins¹, MD, PhD; Pranavsingh Dhunoo^{2,3}, MD, MSc; Márton Görög², MSc; Bertalan Mesko², MD, PhD

¹Department of Surgery, Ikazia Hospital, Rotterdam, Netherlands

²The Medical Futurist Institute, Budapest, Hungary

³Department of Computing, Donegal Campus, Atlantic Technological University, Letterkenny, Ireland

Corresponding Author:

Bertalan Mesko, MD, PhD

The Medical Futurist Institute

Povl Bang-Jensen u 2/B1 4/1

Budapest, 1118

Hungary

Phone: 36 703807260

Email: berci@medicalfuturist.com

Abstract

Background: Artificial intelligence (AI)- and machine learning (ML)-based medical devices and algorithms are rapidly changing the medical field. To provide an insight into the trends in AI and ML in health care, we conducted an international patent analysis.

Objective: It is pivotal to obtain a clear overview on upcoming AI and ML trends in health care to provide regulators with a better position to foresee what technologies they will have to create regulations for, which are not yet available on the market. Therefore, in this study, we provide insights and forecasts into the trends in AI and ML in health care by conducting an international patent analysis.

Methods: A systematic patent analysis, focusing on AI- and ML-based patents in health care, was performed using the Espacenet database (from January 2012 until July 2022). This database includes patents from the China National Intellectual Property Administration, European Patent Office, Japan Patent Office, Korean Intellectual Property Office, and the United States Patent and Trademark Office.

Results: We identified 10,967 patents: 7332 (66.9%) from the China National Intellectual Property Administration, 191 (1.7%) from the European Patent Office, 163 (1.5%) from the Japan Patent Office, 513 (4.7%) from the Korean Intellectual Property Office, and 2768 (25.2%) from the United States Patent and Trademark Office. The number of published patents showed a yearly doubling from 2015 until 2021. Five international companies that had the greatest impact on this increase were Ping An Medical and Healthcare Management Co Ltd with 568 (5.2%) patents, Siemens Healthineers with 273 (2.5%) patents, IBM Corp with 226 (2.1%) patents, Philips Healthcare with 150 (1.4%) patents, and Shanghai United Imaging Healthcare Co Ltd with 144 (1.3%) patents.

Conclusions: This international patent analysis showed a linear increase in patents published by the 5 largest patent offices. An open access database with interactive search options was launched for AI- and ML-based patents in health care.

(JMIR AI 2023;2:e47283) doi:[10.2196/47283](https://doi.org/10.2196/47283)

KEYWORDS

artificial intelligence; patent; healthcare; health care; medical; forecasting; future; AI; machine learning; medical device; open-access; AI technology

Introduction

Artificial intelligence (AI), in the form of machine learning (ML)-based medical devices and algorithms, has been rapidly changing a range of aspects of the medical profession from

clinical decision-making to diagnostic imaging interpretation [1,2]. Both the commercial development and academic research focusing on AI and ML in health care showed an exponential growth; however, regulation for clinical use and commercial rollout follow a slower path.

The US Food and Drug Administration (FDA) has been leading the way for regulators worldwide, being the first regulatory body to adopt an AI policy and provide guidelines for approving AI-based medical technologies in practice [3].

Certain medical specialties stand out in terms of the impact of AI on the practice of those professions. Based on a previous study by our research group, examples include cardiology, radiology, and oncology—medical specialties that entail many data-based tasks and components. A better understanding of which medical specialties will be impacted by AI in the near future might shed light on what guidelines, policies, or frameworks to dedicate enough efforts to next.

Also, while the number of peer-reviewed papers on AI's role in health care and medicine, relevant patents, and commercially available AI and ML devices keeps on growing at an unprecedented rate, it will become increasingly difficult for regulators and policy makers to keep up with the pace of innovation [4].

There are numerous health care- and AI-related patents worldwide. Inventors and researchers can submit their patents to national and international offices, of which the largest ones include the China National Intellectual Property Administration (CNIPA), the European Patent Office (EPO), the Japan Patent Office (JPO), the Korean Intellectual Property Office (KIPO), or the United States Patent and Trademark Office. These 5 largest patent offices collaborate in the Five IP Offices collaboration, making all their patents available in the Global Dossier initiative [5].

Not every patent will lead to a product or a service on the market, and even for those that succeed, it usually takes years to reach the market and end up being a commercially available product or a product used in the medical practice.

For example, a patent was submitted for “wireless transmission of ECGs in handheld devices” in 1998 in the United States [6]. The applicants of the patent developed the idea of a smartphone case that served as a single-lead electrocardiogram to be approved by the FDA in 2012, a total of 14 years later. The evolution of its design resulted in a credit card-sized device and an even smaller version in 2021. In the meantime, the company AliveCor received clearance by the FDA to use an algorithm for the analysis of the readings to determine issues related to cardiac rhythm without human intervention [7]. It took around 2 decades for a digital health technology to transition from a patent phase to becoming commercially available, and years to build AI analysis into the device.

It is pivotal to obtain a clear overview on upcoming AI and ML trends in health care to provide regulators with a better position to foresee what technologies they will have to create regulations for that are not available in the market yet.

Therefore, in this study, we provide insights and forecasts into the trends in AI and ML in health care by conducting an international patent analysis.

Methods

Selection of Patents

We selected the Espacenet search engine of the EPO to access the data from the 5 international patent offices collaborating in the Global Dossier initiative [8-10]. The Global Dossier initiative enables web-based public access for the patent data of the CNIPA [11], EPO, JPO [12], KIPO [13], the United States Patent and Trademark Office [14], and provides computer translations to English for the CNIPA, JPO, and KIPO.

We performed a systematic search for the period between January 1, 2012, and July 20, 2022. A query was made using the following keywords: *deep learning*, *machine learning*, *deep neural networks*, or *artificial intelligence*, in combination with *medical*, *medicine*, *healthcare*, or *health*. In addition, the “computing arrangements based on specific computational models” (G06N) of the Cooperative Patent Classification (CPC) was used [15]. Patents are classified with at least 1 CPC code and the G06N is assigned when the invention relates to AI or ML techniques.

The following variables were extracted from the Espacenet database: patent title and abstract, inventors, applicants, publication number, CPC code, and publication date. As inventors are allowed to submit their patent at multiple patent offices, duplicate patents were removed based on matching titles, inventor names, and applicant names. The patent publication number was used to identify the patent office that registered the patent.

Downloading Patent Abstracts

In total, 12,384 matches were found using the Espacenet search, based on which we performed the analysis. Search queries might contain overlapping results; therefore, we excluded duplications, finally retaining 10,967 distinct matches.

Public information was downloaded for all of the patents using a Chrome-based crawler from Espacenet, followed by the extraction of titles and abstracts from the HTML source. The resulting text data were saved to files for further processing. Crawling was performed in August and September 2022.

Preprocessing of Textual Data

The first publication date and number were used wherever multiple were available.

We retained patents that were dated after January 1, 2016, excluding 197 (1.79%) patents of the available data set. The last fully covered month was June 2022, the few patents (n=27, 0.25%) in July 2022 were excluded.

Some of the most frequent words of the English language were excluded from the analysis, as they would rank high in appearance statistics without highlighting the trends we are looking for. The excluded words were the following: “for,” “from,” “and,” “with,” “on,” “of,” “a,” “the,” “to,” “is,” “an,” “by,” “are,” “in,” “can,” “or,” “that,” and “be.” Additionally, commas and parentheses were removed, and the text was converted to lowercase.

Statistics Generated From Downloaded Data

Multiple statistics were generated from the patents; these were evaluated separately for titles and abstracts as well, except for top lists.

Occurrence Counts: Single Words

Titles of all used patents were merged into 1 string, and the number of times each word appears was counted. The occurrence count list was constructed the same way for abstracts as well. Each word appearance was counted, not limited to 1 per patent.

Abstract Query

Some further cases were not covered by the abovementioned lists: expressions consisting of 2 or more words (eg, “brain ct image”), or words from the abstracts that are not listed above due to a very low occurrence count. A researcher could look up arbitrary texts using the query form.

After performing the preprocessing steps, the query string was looked up in each patent’s title. The number of patents with matches was counted—that is, each patent is counted once at maximum—as opposed to the “occurrence counts” described above.

Additionally, appearance counts were displayed on a time scale as well to visualize trends in 3-month units.

Furthermore, to eliminate the effect of increasing patent count, the relative frequency of the search term was also displayed—this is useful to determine the trends of methods because raw occurrence counts could increase even with a declining technology when total patent counts increase over time.

Top Lists

Inventor, applicant, and CPC top lists are simple lists with occurrence counts, based on patent properties without any pre- or postprocessing steps.

A list of the top 20 medical specialties and related terms was curated ([Multimedia Appendix 1](#)): anesthesiology, cardiology, dentistry, dermatology, emergency medicine, gastroenterology, gerontology, family medicine or primary care, internal medicine (ie, infectiology, endocrinology, and nephrology), neurology, obstetrics and gynecology, oncology, ophthalmology, pathology, pediatrics, psychiatry, pulmonology, radiology or nuclear medicine, surgery, and urology [16].

Open Access, Interactive Database

We made our database open access, which is available on The Medical Futurist website [17]. The page allows visitors to analyze the patent database to validate our findings and discover other trends. The code is available upon request.

Users can select from among the available functions in the left sidebar, while the content for the chosen page appears on the right side.

In this web-based open access database, term frequency–inverse document frequency is applied for the purpose of frequency scoring. Single-word occurrences within titles and abstracts

were introduced above, along with *Query* and *Toplists* pages. Besides these, the most frequent word pairs (eg, image segmentation) are also listed with the number of occurrences separately for titles and abstracts. Finally, under “Trending,” one can find those expressions whose occurrence rises steadily within the last 5 examined quarters, possibly highlighting methods that are currently becoming popular. The “Trending” page examines 3 separate properties: change in absolute and relative occurrence, along with the shape of the increase by quarters correlated to a linearly increasing line in the (the “Trend” column).

Interestingly, most of the single words with a high relative increase are linked to modern technologies within health care (“device,” “forecasting,” “inference,” and “classifying”). Similarly, some of the increasingly used word pairs are “computer aided” and “learning algorithms.”

Results

By using the patent database filter option “Applicant toplist,” a list in descending order of the number of patent applications per applicant was generated. The number of patents applied by an entity ranged from 1 to 305, with applicant Ping An Technology (Shenzhen, China) filing for the highest number of patents (n=305) and several dozens of applicants filing for the lowest number of patents (n=1). We identified 5848 patents with a company as the primary applicant and 3038 patents with a university as the primary applicant. To derive insights relevant for the purposes of this study, the 20 applicants from this list, which applied for the most patents, were considered and the findings are summarized in [Table 1](#).

Each entry in the “Applicant toplist” filter also lists the corresponding country, in abbreviated format, where the relevant patent office is located. Out of the top 20 patent applicants, 14 are based in China, 3 are based in the United States, and 1 is based in Germany, Japan, and the Netherlands, each.

From these data and extending to applicants beyond the top 20 ones, a list of the top 10 countries where most patents were applied from was curated. As [Table 1](#) indicates, most of the relevant patent applications were filed in China, followed by the United States. Among this list of top 10 countries, 4 are located in Asia, 4 are located in Europe, and 2 are located in North America. [Textbox 1](#) shows the top 10 countries from where relevant patents were filed.

By selecting the “Patent office stats” option from the database, the general trend in the number of health care patents between 2016 and 2022 in selected patent offices was observed. There were 156, 340, 747, 1552, 253, 4097, and 1278 AI- and ML-related health care patents in 2016, 2017, 2018, 2019, 2020, and July 2021, respectively; this indicates a general increase in the application of such patents during that time period in the patent offices in China, the United States, and South Korea, while the offices in Japan and Spain have experienced little to no change in the volume of patents. [Figure 1](#), generated from the database, plots the number of patents in the selected patent offices over this time period.

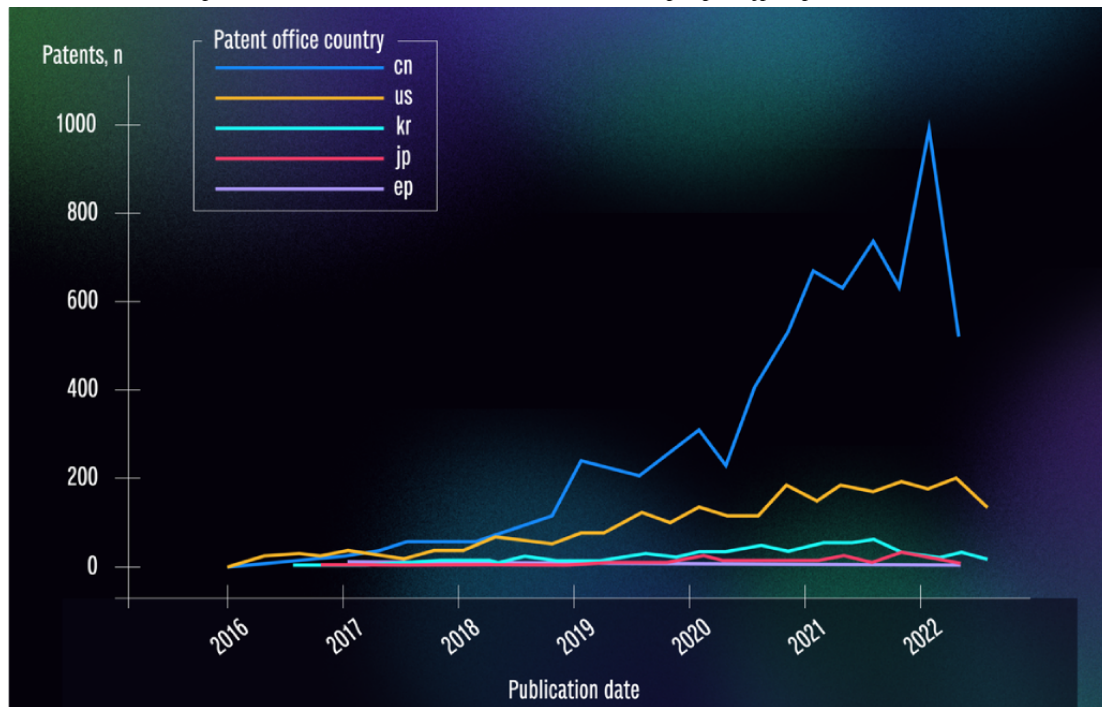
Table 1. Top 20 patent applicants.

Number	Applicant name	Occurrences, n	Country
1	Ping An Technology (Shenzhen) Co Ltd	305	China
2	Siemens Healthcare GmbH	219	Germany
3	IBM Corp	217	United States
4	Koninklijke Philips N.V.	110	The Netherlands
5	Ping An Medical and Healthcare Management Co Ltd	105	China
6	Ping An International Smart City Technology Co Ltd	103	China
7	Tencent Technology Shenzhen Co Ltd	90	China
8	University of Electronic Science and Technology of China	82	China
9	Zhejiang University	79	China
10	Shandong University	59	China
11	Beijing University of Technology	57	China
12	Tsinghua University	57	China
13	Fudan University	50	China
14	Canon Medical Systems Corporation	47	Japan
15	Beijing Baidu Netcom Science Technology Co Ltd	46	China
16	Tianjin University	45	China
17	GE Precision Healthcare LLC	45	United States
18	Huazhong University of Science and Technology	45	China
19	Beihang University	44	China
20	General Electric	44	United States

Textbox 1. Top 10 countries from where patents were filed.

<p>The top 10 countries from where patents were filed were as follows:</p> <ol style="list-style-type: none"> 1. China 2. United States 3. South Korea 4. Germany 5. Japan 6. The Netherlands 7. Canada 8. India 9. United Kingdom 10. France

Figure 1. Patent trends in selected patent offices between 2016 and 2022. cn: China; ep: Spain; jp: Japan; kr: South Korea; us: United States.



The rate of increase in the number of patents varied in each office that experienced such an increase. A marked increase was noticeable in China from mid-2017, around 2018 in the United States, and only around 2020 in South Korea. The patent office in China experienced a steady increase in the number of applications with some notable dips in 2020, 2021, and 2022. Despite those downtimes, that patent office maintained its lead during the time period analyzed.

To analyze patent trends around medical specialties, we created a database of words and expressions that are relevant to each of the major 20 medical specialties (Multimedia Appendix 1).

When analyzing single words that appear in the title of patents, the top 5 medical specialties with the highest number of patents were radiology, oncology, cardiology, pulmonology, and surgery with 394, 271, 128, 103, and 76 patents, respectively.

The “Abstract - query” option of the database outputs the number of times the search term occurs in the abstracts. Using the preselected specific terms for medical specialties, the occurrence of specialty-related terms was identified. Based on this list, the terms relating to radiology or nuclear medicine occurred the most in abstracts (n=1160), followed by oncology (n=532), ophthalmology (n=454), surgery (n=309), pulmonology (n=261), cardiology (n=252), and obstetrics and gynecology (n=217; Table 2).

When focusing on one of the medical specialties with a high number of patents (for instance, radiology), trends in imaging-based patents could be established. The 8 most frequently used imaging-related terms were “image processing” (n=682), “image data” (n=674), “imaging” (n=657), “image segmentation” (n=328), “CT image” (n=288), “X-ray” (n=120), “MRI” (n=114), and “ultrasound” (n=77). An increase in the occurrence of these imaging-related terms was identified between 2015 and 2021 (Figure 2). For the field of oncology,

trends showed a similar increase. The 4 most used terms were “cancer” (n=161), “tumor” (n=151), “radiotherapy” (n=55), and “malignant” (n=47).

When focusing on terms related to AI and ML, trends in AI- and ML-based patents could be established. The 4 most used AI- and ML-based terms were “artificial intelligence” (n=2450), “neural network” (n=2043), “machine learning” (n=1717), and “deep learning” (n=1492). An increase in the occurrence of these AI- and ML-based terms was identified between 2015 and 2021 (Figure 3).

To demonstrate what kind of patents were included in the database, we chose to feature examples of recently registered patents of the top 4 applicants: Ping An Group listed a patent within the scope of the specialties of radiology and oncology, titled “Lymph node metastasis prediction method and device, equipment and storage medium” (CN113920137a) in January 2022. This patent focuses on the detection of lymph node metastasis in pancreatic ductal cancer on computed tomographic imaging of the abdomen. The results of the first clinical application were published in January 2023 [18].

Siemens Healthineers AG listed a patent within the scope of the specialties of radiology and pulmonology, titled “Assessment of abnormality patterns associated with covid-19 from x-ray images” (US2022022818a) in January 2022. A full package of AI solutions for COVID-19 imaging became commercially available the months thereafter [19].

IBM Corp listed a patent within the scope of the specialties of pathology and oncology, titled “Interpretation of whole-slide images in digital pathology” (US2022164946A1) in May 2022. The code, data, and models were published in January 2022 and a Python-based package (for modeling and learning) is freely available on GitHub [20,21].

Koninklijke Philips N.V. listed a patent within the scope of the specialty of cardiology, titled “Systems and methods for identifying low clinical value telemetry cases” (US2022020478A1) in January 2022. This patent is part of the

Philips Cardiologs arrhythmias diagnostic software, which is commercially available and FDA-cleared under section 510(k) of the Food, Drug and Cosmetic Act [22].

Table 2. Occurrence of specialty-related terms.

Number	Specialty	Occurrences, n
1	Anesthesiology	72
2	Dentistry	41
3	Cardiology	252
4	Dermatology	112
5	Emergency medicine	157
6	Gastroenterology	84
7	Gerontology	37
8	Family medicine or primary care	28
9	Internal medicine	174
10	Neurology	77
11	Obstetrics and gynecology	217
12	Oncology (ie, radiation oncology)	532
13	Ophthalmology	454
14	Pathology	87
15	Pediatrics	18
16	Psychiatry	94
17	Pulmonology	261
18	Radiology	1160
19	Surgery	309
20	Urology	28

Figure 2. Trends in imaging-based patents. CT: computed tomography; MRI: magnetic resonance imaging.

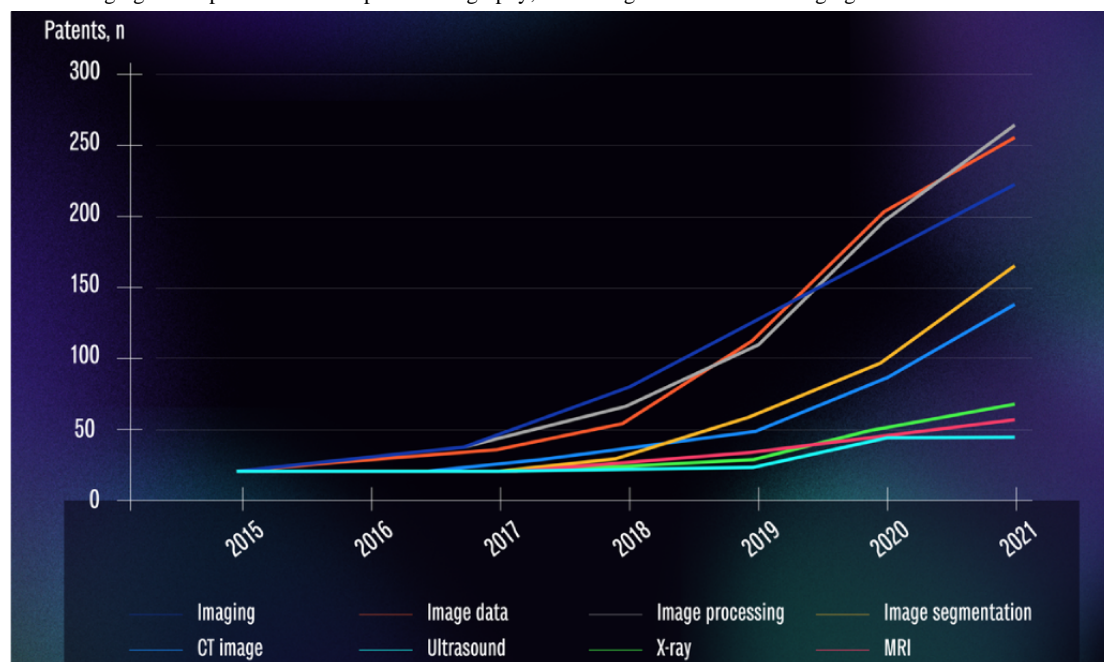
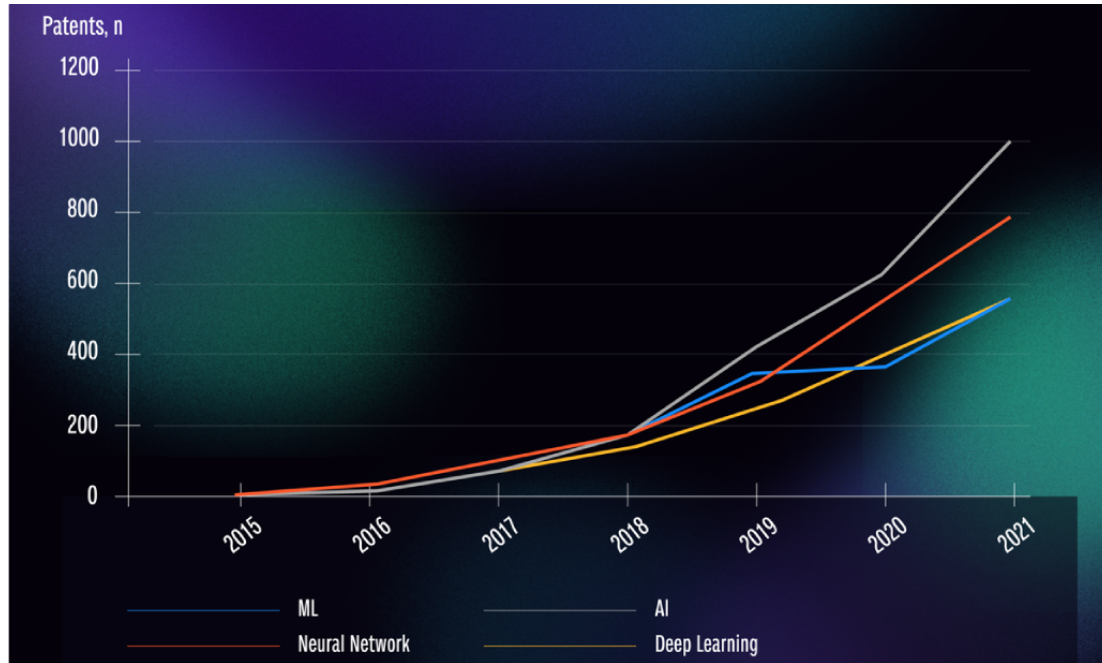


Figure 3. Trends in AI- and ML-based patents. AI: artificial intelligence; ML: machine learning.



Discussion

Principal Findings

Based on the identified health care-related AI-based patents, China clearly stands out as a leader in the field of AI. Also, 2020 seems to be a turning point with marked growth in health care-related patents. Following the widespread success of ChatGPT (OpenAI) in 2022, there are no data to indicate that this growth would slow down [23].

Certain medical specialties stand out in terms of the number of patents that have been submitted about AI technologies and inventions that might be relevant to them (Figure 4). Based on a previous study published by our group about FDA-approved AI- and ML-based medical technologies [1], radiology, cardiology, and oncology were already identified as specialties with many AI-based applications. The more repetitive or data-based tasks a specialty entails, the higher the potential for automation to be able to contribute to that field.

Moreover, patents that include the analysis of medical images or videos can be relevant to a range of specialties from radiology to pulmonology and surgery. Specialties that are closely linked to medical imaging can also be in the focus of AI patents in the coming years. Examples include dentistry, ophthalmology, and emergency medicine.

Besides these imaging-oriented specialties, as analyzing images is a widely popular use case of AI and ML, dermatology and pathology could also benefit from the AI revolution. In dermatology, the rise of skin-checking applications that can analyze photos of skin lesions on patients' smartphones underscores this observation [24]. In pathology, automated assessment of digitized histopathology slides falls into the same category [25].

Medical specialties such as psychiatry or neurology that are more interaction-based (as opposed to being data-based) and

entail more creative (vs repetitive) tasks might receive fewer AI patents; thus, those could be less prone to AI- or ML-based innovations [26].

The discrepancy between the top-ranking medical specialties in the title and abstract analyses could be attributed to the higher occurrence of related terms in the abstracts than in the titles, given the higher density of words in the former.

With this study, we attempt to prove the point that in the age of automation, preparing with regulations in time should be of high priority among policy makers. The #wearenotwaiting movement that comprises thousands of patients with diabetes, who created artificial pancreatic systems, further emphasizes this [27]. These patients have developed applications, platforms, and other solutions to help each other manage their diabetes. Their OpenAPS (Open Artificial Pancreas System) software that was created entirely by the patient community with no contribution from medical professionals automatically provides patients with the right doses of insulin based on their blood glucose level [28].

Due to the influx of advanced technologies such as wearable health sensors, portable diagnostic devices, and AI and ML applications in health care, it has become inevitable to design regulations and guidelines for technologies that are not available in the market yet, but everything, including patent trends, indicates that they will soon be. As patients now have access to technologies, data, and algorithms, they will find a way to use the technology that is not yet regulated but can still help them manage their condition or health.

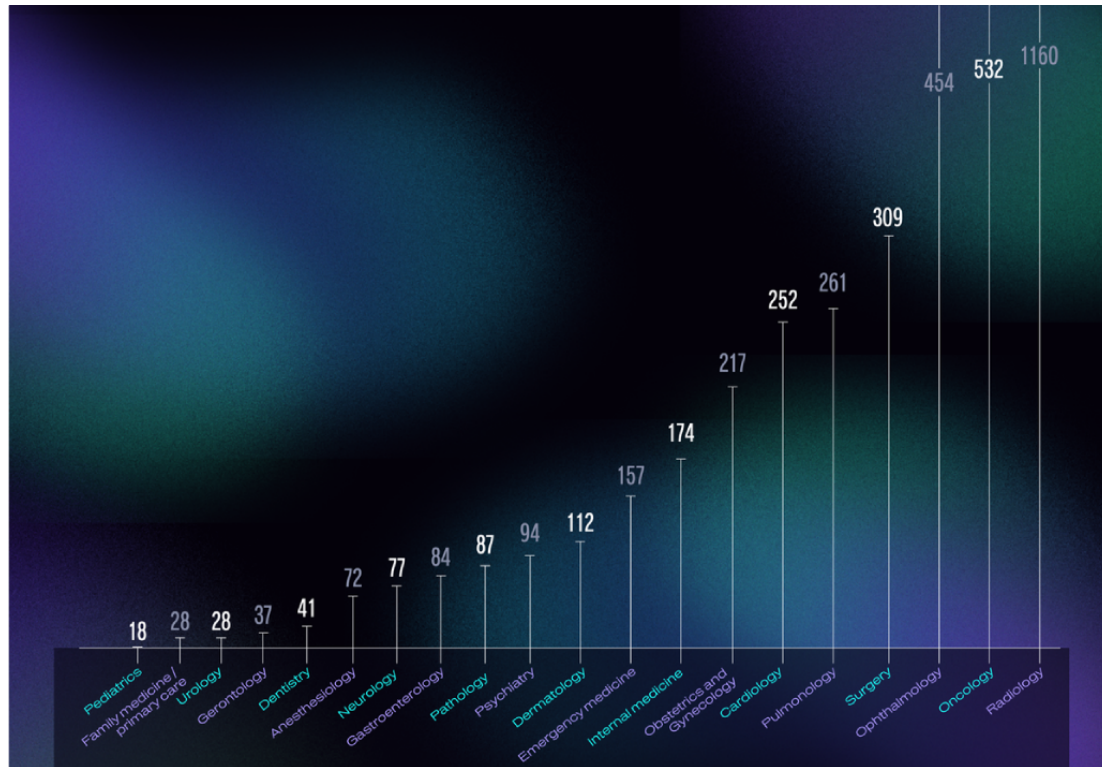
The recent rise of the conversational agent and large language model ChatGPT and AI-based image generators such as MidJourney and DALL-E all point toward this direction. As a response to ChatGPT, Google LLC published a study about their own chatbot that was specifically designed to answer medical questions [29].

We expect that by looking at medical and health care–related AI and ML patent trends, regulators and policy makers could better determine medical specialties, technological trends, or areas such as imaging to dedicate more attention to. Thus, when a range of AI- and ML-based technologies become available in those fields, proper regulations will ensure a safe and efficient

implementation into the practice of medicine and the delivery of health care.

A follow-up study that closely follows some of the patents and medical specialties that stood out in this analysis would be useful to see and determine how much time it takes for an AI- or ML-based health care patent to reach the stage of practical implementation.

Figure 4. The number of occurrences of specialty-related terms in healthcare AI patents assigned to each of the 20 medical specialties. AI: artificial intelligence.



Limitations

There are obvious limitations to our approach. As there is no globally accepted patent database, we could only focus on the 5 most active patent offices with the highest number of patents worldwide. This implies that we might have overlooked patents from other patent offices worldwide. As there is no database in

the literature about what keywords and expressions might be associated with certain medical specialties, the database we generated is a subjective list of keyword-specialty associations. Moreover, even if a specific medical specialty or its keyword is mentioned in a patent's abstract, it does not necessarily mean that the patents are indeed associated with the specialty.

Authors' Contributions

SB, PD, and BM designed and conducted the study, GM designed the database, and all authors wrote the manuscript.

Conflicts of Interest

None declared.

Multimedia Appendix 1

Medical specialties and related terms.

[PDF File (Adobe PDF File), 92 KB - [ai_v2i1e47283_app1.pdf](#)]

References

1. Benjamins S, Dhunoo P, Meskó B. The state of artificial intelligence-based FDA-approved medical devices and algorithms: an online database. *NPJ Digit Med* 2020;3:118 [FREE Full text] [doi: [10.1038/s41746-020-00324-0](https://doi.org/10.1038/s41746-020-00324-0)] [Medline: [32984550](https://pubmed.ncbi.nlm.nih.gov/32984550/)]
2. Topol E. High-performance medicine: the convergence of human and artificial intelligence. *Nat Med* 2019 Jan;25(1):44-56. [doi: [10.1038/s41591-018-0300-7](https://doi.org/10.1038/s41591-018-0300-7)] [Medline: [30617339](https://pubmed.ncbi.nlm.nih.gov/30617339/)]

3. Artificial intelligence and machine learning in software as a medical device. US Food and Drug Administration. 2021. URL: <https://tinyurl.com/rwrh739a> [accessed 2023-05-10]
4. Buiten M. Towards intelligent regulation of artificial intelligence. *Eur J Risk Regul* 2019 Apr 29;10(1):41-59 [FREE Full text] [doi: [10.1017/err.2019.8](https://doi.org/10.1017/err.2019.8)]
5. Five IP Offices. URL: <https://www.fiveipoffices.org/home> [accessed 2023-05-10]
6. Buntz B. Ahead of his time. *Medical Device and Diagnostic Industry*. 2012. URL: <https://www.mddionline.com/news/ahead-his-time> [accessed 2023-05-10]
7. Moynihan T. Ex-Googlers build a neural network to protect your heart. *Wired*. 2017. URL: <https://www.wired.com/2017/03/alivecor-kardia/> [accessed 2023-05-10]
8. Espacenet: free access to over 140 million patent documents. European Patent Office. URL: <https://worldwide.espacenet.com/patent/> [accessed 2023-05-10]
9. European Patent Office. URL: <https://www.epo.org/> [accessed 2023-05-10]
10. Global Dossier and patent information (Working Group 2). Five IP Offices. URL: <https://www.fiveipoffices.org/activities/globaldossier> [accessed 2023-05-10]
11. China National Intellectual Property Administration. URL: <https://english.cnipa.gov.cn/> [accessed 2023-05-10]
12. Japan Patent Office. URL: <https://www.jpo.go.jp/e/index.html> [accessed 2023-05-10]
13. Korean Intellectual Property Office. URL: <https://www.kipo.go.kr/en/> [accessed 2023-05-10]
14. United States Patent and Trademark Office. URL: <https://www.uspto.gov/> [accessed 2023-05-10]
15. CPC scheme and CPC definitions. Cooperative Patent Classification. URL: <https://www.cooperativepatentclassification.org/cpcSchemeAndDefinitions/table> [accessed 2023-05-10]
16. Association of American Medical Colleges. URL: <https://www.aamc.org> [accessed 2023-05-10]
17. The Medical Futurist. URL: <http://www.medicalfuturist.com/patents> [accessed 2023-05-11]
18. Bian Y, Zheng Z, Fang X, Jiang H, Zhu M, Yu J, et al. Artificial intelligence to predict lymph node metastasis at CT in pancreatic ductal adenocarcinoma. *Radiology* 2023 Jan;306(1):160-169. [doi: [10.1148/radiol.220329](https://doi.org/10.1148/radiol.220329)] [Medline: [36066369](https://pubmed.ncbi.nlm.nih.gov/36066369/)]
19. AI COVID-19. Siemens Healthcare GmbH. URL: <https://www.siemens-healthineers.com/medical-imaging/digital-transformation-of-radiology/ai-covid-19-algorithm> [accessed 2023-05-10]
20. Hierarchical Graph Representations in Digital Pathology. GitHub. URL: <https://github.com/histocartography/hact-net> [accessed 2023-05-10]
21. Gibbs S. Artificial intelligence tool 'as good as experts' at detecting eye problems. *The Guardian*. 2018. URL: <https://www.moorfields.nhs.uk/sites/default/files/uploads/documents/Final> [accessed 2022-11-09]
22. AI-enabled solutions. Philips. URL: https://www.philips.com/a-w/about/artificial-intelligence/ai-enabled-solutions#triggername=less33_isc33 [accessed 2023-05-10]
23. van Dis EAM, Bollen J, Zuidema W, van Rooij R, Bockting CL. ChatGPT: five priorities for research. *Nature* 2023 Feb;614(7947):224-226 [FREE Full text] [doi: [10.1038/d41586-023-00288-7](https://doi.org/10.1038/d41586-023-00288-7)] [Medline: [36737653](https://pubmed.ncbi.nlm.nih.gov/36737653/)]
24. Novoa R, Gevaert O, Ko J. Marking the path toward artificial intelligence-based image classification in dermatology. *JAMA Dermatol* 2019 Oct 01;155(10):1105-1106 [FREE Full text] [doi: [10.1001/jamadermatol.2019.1633](https://doi.org/10.1001/jamadermatol.2019.1633)] [Medline: [31411643](https://pubmed.ncbi.nlm.nih.gov/31411643/)]
25. Niazi M, Parwani AV, Gurcan M. Digital pathology and artificial intelligence. *Lancet Oncol* 2019 May;20(5):e253-e261 [FREE Full text] [doi: [10.1016/S1470-2045\(19\)30154-8](https://doi.org/10.1016/S1470-2045(19)30154-8)] [Medline: [31044723](https://pubmed.ncbi.nlm.nih.gov/31044723/)]
26. Doraiswamy P, Blease C, Bodner K. Artificial intelligence and the future of psychiatry: insights from a global physician survey. *Artif Intell Med* 2020 Jan;102:101753 [FREE Full text] [doi: [10.1016/j.artmed.2019.101753](https://doi.org/10.1016/j.artmed.2019.101753)] [Medline: [31980092](https://pubmed.ncbi.nlm.nih.gov/31980092/)]
27. Wilmot E, Danne T. DIY artificial pancreas systems: the clinician perspective. *Lancet Diabetes Endocrinol* 2020 Mar;8(3):183-185 [FREE Full text] [doi: [10.1016/s2213-8587\(19\)30416-4](https://doi.org/10.1016/s2213-8587(19)30416-4)]
28. Lewis D. Do-it-yourself artificial pancreas system and the OpenAPS movement. *Endocrinol Metab Clin North Am* 2020 Mar;49(1):203-213 [FREE Full text] [doi: [10.1016/j.ecl.2019.10.005](https://doi.org/10.1016/j.ecl.2019.10.005)] [Medline: [31980119](https://pubmed.ncbi.nlm.nih.gov/31980119/)]
29. Abduljawad M, Alsalmami A. Towards creating exotic remote sensing datasets using image generating AI. 2022 Presented at: 2022 International Conference on Electrical and Computing Technologies and Applications (ICECTA); November 23-25, 2022; Ras Al Khaimah, United Arab Emirates. [doi: [10.1109/icecta57148.2022.9990245](https://doi.org/10.1109/icecta57148.2022.9990245)]

Abbreviations

- AI:** artificial intelligence
- CNIPA:** China National Intellectual Property Administration
- CPC:** Cooperative Patent Classification
- EPO:** European Patent Office
- FDA:** US Food and Drug Administration
- JPO:** Japan Patent Office
- KIPO:** Korean Intellectual Property Office
- ML:** machine learning

Edited by K El Emam; submitted 14.03.23; peer-reviewed by A Gao, JH Rajendran; comments to author 15.04.23; revised version received 05.05.23; accepted 07.05.23; published 26.05.23.

Please cite as:

Benjamins S, Dhunnoo P, Görög M, Mesko B

Forecasting Artificial Intelligence Trends in Health Care: Systematic International Patent Analysis

JMIR AI 2023;2:e47283

URL: <https://ai.jmir.org/2023/1/e47283>

doi: [10.2196/47283](https://doi.org/10.2196/47283)

PMID: [10449890](https://pubmed.ncbi.nlm.nih.gov/10449890/)

©Stan Benjamins, Pranavsingh Dhunnoo, Márton Görög, Bertalan Mesko. Originally published in JMIR AI (<https://ai.jmir.org>), 26.05.2023. This is an open-access article distributed under the terms of the Creative Commons Attribution License (<https://creativecommons.org/licenses/by/4.0/>), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR AI, is properly cited. The complete bibliographic information, a link to the original publication on <https://www.ai.jmir.org/>, as well as this copyright and license information must be included.

Review

Application of a Comprehensive Evaluation Framework to COVID-19 Studies: Systematic Review of Translational Aspects of Artificial Intelligence in Health Care

Aaron Edward Casey^{1,2}, PhD; Saba Ansari³, GCHE (Teaching and Learning), MSc; Bahareh Nakisa⁴, BSE, MCS, PhD; Blair Kelly⁵, Grad Dip (InfoLibStds), BCom; Pieta Brown⁶, MPS; Paul Cooper³, PhD; Imran Muhammad³, MIS, MSc, PhD; Steven Livingstone⁶, BSc, GradDipSci, MDataSci; Sandeep Reddy³, MBBS, MSc, PhD; Ville-Petteri Makinen^{1,2,7,8}, DSc

¹South Australian Health and Medical Research Institute, Adelaide, Australia

²Australian Centre for Precision Health, Cancer Research Institute, University of South Australia, Adelaide, Australia

³School of Medicine, Deakin University, Geelong, Australia

⁴School of Information Technology, Deakin University, Geelong, Australia

⁵Library, Deakin University, Geelong, Australia

⁶Orion Health, Auckland, New Zealand

⁷Computational Medicine, Faculty of Medicine, University of Oulu, Oulu, Finland

⁸Centre for Life Course Health Research, Faculty of Medicine, University of Oulu, Oulu, Finland

Corresponding Author:

Aaron Edward Casey, PhD

South Australian Health and Medical Research Institute

North Terrace

Adelaide, 5000

Australia

Phone: 61 08 8128 4064

Email: aaron.casey@sahmri.com

Abstract

Background: Despite immense progress in artificial intelligence (AI) models, there has been limited deployment in health care environments. The gap between potential and actual AI applications is likely due to the lack of translatability between controlled research environments (where these models are developed) and clinical environments for which the AI tools are ultimately intended.

Objective: We previously developed the Translational Evaluation of Healthcare AI (TEHAI) framework to assess the translational value of AI models and to support successful transition to health care environments. In this study, we applied the TEHAI framework to the COVID-19 literature in order to assess how well translational topics are covered.

Methods: A systematic literature search for COVID-19 AI studies published between December 2019 and December 2020 resulted in 3830 records. A subset of 102 (2.7%) papers that passed the inclusion criteria was sampled for full review. The papers were assessed for translational value and descriptive data collected by 9 reviewers (each study was assessed by 2 reviewers). Evaluation scores and extracted data were compared by a third reviewer for resolution of discrepancies. The review process was conducted on the Covidence software platform.

Results: We observed a significant trend for studies to attain high scores for technical capability but low scores for the areas essential for clinical translatability. Specific questions regarding external model validation, safety, nonmaleficence, and service adoption received failed scores in most studies.

Conclusions: Using TEHAI, we identified notable gaps in how well translational topics of AI models are covered in the COVID-19 clinical sphere. These gaps in areas crucial for clinical translatability could, and should, be considered already at the model development stage to increase translatability into real COVID-19 health care environments.

(JMIR AI 2023;2:e42313) doi:[10.2196/42313](https://doi.org/10.2196/42313)

KEYWORDS

artificial intelligence; health care; clinical translation; translational value; evaluation; capability; utility; adoption; COVID-19; AI application; health care AI; model validation; AI model; AI tools

Introduction

The discussion about the value of artificial intelligence (AI) to health care and how AI can address health care delivery issues has been in place for some years now [1-3]. However, most stakeholders are eager for this discourse to move beyond theoretical or experimental confines to adoption and integration in clinical and real-world health care environments [1,4,5]. Recently, we have started to see some AI applications undergoing clinical trials or integration into medical devices or medical information systems [6]. Yet, most AI applications in health care have not demonstrated improvement in clinical or health care outcomes [5,7]. What prevents these applications from translating their potential to clinical outcomes? First, many of these AI applications are developed to demonstrate algorithmic performance or superiority rather than improvement in clinical results [8,9]. Second, the applications are not considered for use beyond the experimental or pilot settings [8]. This limitation means their performance does not often generalize beyond test data sets. Third, even when these applications are externally validated, they are seldom integrated into existing clinical workflows, often because of decreased performance on the external validation [10] or low acceptance by clinicians [11]. The latter aspect means these applications remain experimental novelties rather than useful tools for clinicians. Added to these translational issues are problems with data that may lead to inaccurate results or the introduction of biases. Several studies have shown how such issues can have adverse outcomes for patients and communities [12-14]. Yet, ethical and governance safeguards are often missing in AI in health care applications or studies [14].

These translational issues suggest there is a need for a comprehensive framework that can support researchers, software vendors, and relevant parties in systematically assessing their AI applications for their translational potential. To address this gap, we formed an international team and ran a systematic

process over 18 months to develop an evaluation and guidance framework, termed “Translational Evaluation of Healthcare AI” (TEHAI) [15]. This framework focuses on the aspects that can support the practical implementation and use of AI applications. TEHAI has 3 main domains (capability, utility, and adoption components) and 15 subcomponents (Table 1 and Multimedia Appendix 1). As the range of clinical challenges and potential AI solutions is wide, it is infeasible to automate the evaluation using current technology. Instead, we rely on TEHAI as an expert-driven but formalized framework where the subjectivity of an individual reviewer is mitigated by the consensus power of multiple committee members.

The emergence of the COVID-19 pandemic has resulted in several studies and papers outlining the utility of AI in tackling various aspects of the disease, such as diagnosis, treatment, and surveillance [16-19]. The number of AI papers published either as preprints or as peer-reviewed papers has been unprecedented, even leading to the development of AI applications to keep up with and summarize the findings for scientists [20]. Some recent reviews have outlined how most of these studies or the AI applications presented in these studies have shown minimal value for clinical care [7,21]. This finding aligns with the discussion about the translational problem of AI in health care.

The aim of this study is to assess the awareness and consideration for important translational factors in the scientific literature related to COVID-19 machine learning applications. We chose the narrow scope to ensure that our method of evaluation (ie, TEHAI) would not be confounded by the differences that are inherent to any particular area of health care. For this reason, we included only studies where AI was clearly aimed at solving a practical problem rather than discovering new biology or novel treatments. This cost-effective approach enabled us to uncover translational gaps in the AI applications and validate the usefulness of a variety of AI models without the added complexity due to a high diversity of diseases or health care challenges.

Table 1. Overview of the TEHAI^a framework^b.

Component and subcomponents	Initial score	Weight
Capability		
Objective of the study	0-3	10
Data set source and integrity	0-3	10
Internal validity	0-3	10
External validity	0-3	10
Performance metrics	0-3	10
Use case	0-3	5
Utility		
Generalizability and contextualization	0-3	10
Safety and quality	0-3	10
Transparency	0-3	10
Privacy	0-3	10
Nonmaleficence	0-3	10
Adoption		
Use in a health care setting	0-3	10
Technical integration	0-3	10
Number of services	0-3	5
Alignment with the domain	0-3	5

^aTEHAI: Translational Evaluation of Healthcare Artificial Intelligence (AI).

^bThe framework comprises 15 separate criteria (subcomponents) that are grouped into 3 higher-level components. Each criterion yields a score between 0 and 3 points, depending on the quality of the study. To compare 2 or more AI models against each other, further weighting of the scores can be applied to emphasize translatability. However, in this study, weighting was not used, since we focused on the statistics of the subcomponents instead.

Methods

Data Extraction

Eligible studies included those where a statistical algorithm was applied to or trained with a COVID-19 data set and where the intended use of the algorithm was to address a COVID-19 health care problem. Excluded studies included those where participants were younger than 18 years and where the full text of the study was not in English. To find papers eligible for this study, we searched the National Institutes of Health (NIH) iSearch COVID-19 portfolio, MEDLINE via Ovid, and Embase. These sources were searched on December 7, 2020, using search strategies consisting of keywords expected to appear in the title or abstract of eligible studies and index terms specific to each database except in the case of the NIH iSearch COVID-19 portfolio. The search strategy was developed by a health librarian (author BK) in consultation with the rest of the research team.

For the COVID-19 element of the search, we adapted the Wolters Kluwer expert search for COVID-19 on MEDLINE. Specifically, we removed the search lines for excluding non-COVID-19 coronaviruses (eg, Middle East respiratory syndrome) and for pharmaceutical treatment options (eg, remdesivir); at the time our search strategy was created, these were lines 5 and 9, respectively, in the Wolters Kluwer Ovid COVID-19 expert search. For the AI element of the search, we

searched MEDLINE for relevant papers, recording significant keywords from their titles and abstracts. We also searched the Medical Subject Headings (MeSH) thesaurus for related MeSH terms. These steps led to the creation of a draft search strategy, which was then tested and finalized. The search was limited to records with a publication date of December 1, 2019, onward. This limit was to reduce the number of irrelevant results, given that the first known case of COVID-19 occurred in December 2019 ([Multimedia Appendix 2](#)).

A foundational Ovid MEDLINE search strategy was then translated for Embase to make use of appropriate syntax and index terms ([Multimedia Appendix 2](#)). Similar translation was done for the NIH iSearch COVID-19 portfolio except for index terms as this resource did not use indexing at the time of search development ([Multimedia Appendix 2](#)). Finally, search strategy validation and refinement took place by testing a set of known relevant papers against the search strategy, as developed, with all papers subsequently recalled by the search in MEDLINE and Embase. A full reproduction of the search strategies for each database can be found in [Multimedia Appendix 2](#). Searching these databases using the search strategy resulted in 5276 records. After removal of duplicates, we screened 3830 (72.6%) records for relevance. This resulted in 968 (25.3%) studies identified as relevant and eligible for evaluation. From these, a sample of 123 (12.7%) was randomly selected for evaluation and data extraction, of which 102 (82.9%) were included in the final set. Our target number for full evaluation

was 100; however, additional papers were randomly picked to account for the rejection of 21 (17.1%) papers that passed the initial screen but were deemed ineligible after closer inspection (Multimedia Appendix 3). Early on in the evaluation, it became apparent that a significant portion of the studies focused on image analysis; we then enriched the pool for studies that were not imaging focused, taking the ratio of imaging-focused:nonimaging-focused studies to 1:1. The full text was retrieved for all 123 (12.7%) studies in the randomized sample; however, only 102 (82.9%) studies met our inclusion criteria at the evaluation and extraction stage (Multimedia Appendix 4). Of the studies that did not meet our inclusion criteria, the majority were nonimaging studies and the final ratio of imaging-focused:nonimaging-focused studies was 2:1.

Evaluation and data extraction were conducted using Covidence systematic review software [22]. We used this software to facilitate the creation of a quality assessment template based on the TEHAI framework [15] in combination with other questions (henceforth referred to as data extraction questions) aimed at further understanding the components that may influence a study's capacity to translate into clinical practice (Multimedia Appendix 1). As a measure to minimize the impact of subjectivity introduced by human evaluation, each paper was initially scored by 2 reviewers, who independently evaluated the paper against the elements of the TEHAI framework and extracted relevant data. A third reviewer then checked the scores, and if discrepancies were present, they chose 1 of the 2 independent reviewers' scores as the final result. This process was built-in to the Covidence platform. To further minimize the impact of subjectivity introduced by human evaluation, reviewer roles were also randomly assigned across the evaluation team.

For scoring of the included studies, we derived upon previously provided guidance for scoring evidence within the TEHAI framework [15]. The TEHAI framework is composed of 3 overarching components: capability, utility, and adoption. Each component comprises numerous subcomponent questions, of which there are 15 in total. The scoring of each TEHAI subcomponent is based on a range of 0-3, depending on the criteria met by the study. In this study, we also investigated the sums of these scores at the component level to provide a better overview of data. In addition, TEHAI facilitates direct comparisons between specific studies using a weighting mechanism that further emphasizes the importance of translatability (see the last column in Table 1). However, for this study, where we focused on the aggregate statistical patterns, weighting was not used.

We also asked reviewers to report on a select number of data extraction questions that would enable us to further tease apart which components of a study may influence the score obtained. These questions covered (1) the broad type of the AI algorithm, (2) methodological or clinical focus, (3) open source or proprietary software, (4) the data set size, (5) the country of origin, and (6) imaging or nonimaging data.

Data Analysis

Associations between groupings of papers and the distributions of subcomponent scores were assessed with the Fisher exact

test. Correlations between subcomponents were calculated using the Kendall formula. Component scores were calculated by adding the relevant subcomponent scores together; group differences in mean component scores were assessed using the t-test. As there are 15 subcomponents, we set a multiple testing threshold of $P < .003$ to indicate 5% type 1 error probability under the Bonferroni correction for 15 independent tests. Unless otherwise indicated, mean (SE) scores were calculated.

Results

TEHAI Subcomponent Scores

A total of 102 manuscripts were reviewed by 9 reviewers (mean 22.67 per reviewer, SD 7.71, min.=11, max.=36), with the same 2 reviewers scoring the same manuscript an average of 2.83 times (SD 2.58, min.=0, max.=13). The Cohen κ statistic for interreviewer reliability was 0.45, with an asymptomatic SE of 0.017 over the 2 independent reviewers. The reviewer scores were in moderate agreement ($\kappa=0.45$) according to Cohen's original tiers [23]. In practice, this means that the scoring system was successful in capturing important and consistent information from the COVID-19 papers, but there would be too much disagreement due to reviewer background or random noise for demanding applications, such as clinical diagnoses [24]. Given that the role of the TEHAI framework is to provide guidance and decision support (not diagnoses), moderate accuracy is sufficient for a meaningful practical benefit for AI development. Nevertheless, the question of reviewer bias should be revisited in future updates to the framework.

Overall, the capability component scored the highest mean score, followed by adoption and utility (Figure 1A). At the subcomponent level, the poorest-performing questions were nonmaleficence (93/102, 91.2%, scoring 0 points), followed closely by safety and quality, external validity, and the number of services (Figure 1B).

We observed moderate positive correlation ($R=0.19-0.43$) between most capability component questions (data source vs: internal validation $R=0.43$, external validation $R=0.20$, performance $R=0.33$, and use case $R=0.37$; internal validation vs: performance $R=0.40$, use case $R=0.31$; performance vs use case $R=0.32$), with the exception of the subcomponent objective of study (objective of study vs: data source $R=0.13$, internal validation $R=0.09$, external validation $R=0.08$); see Figure 2. This indicated that if a study scored well in one subcomponent of the capability component, then it was also likely to score well in the other capability subcomponents, with the exception of the "objective of the study" subcomponent. Furthermore, there was also a correlation between the subcomponents belonging to the capability component and the "generalizability and contextualization" ($R=0.19-0.31$), "transparency" ($R=0.11-0.27$), and "alignment with the domain" ($R=0.13-0.40$) subcomponents, as well as our data extraction question 9 (method of machine learning used; $R=0.11-0.24$); see Figure 2. There was also a significant, moderate correlation between most adoption component questions ($R=0.18-0.42$), with the exception of the "alignment with the domain" subcomponent ($R=0.04-0.26$); see Figure 2. A significant negative correlation was observed between a country's gross domestic product

(GDP) and imaging studies ($R=-0.30$), indicating that high-GDP countries are less likely to conduct imaging studies than middle-GDP countries. The negative correlation between the audience (clinical or methodological) and the number of services ($R=-0.36$) indicated that methodological studies are less likely

to be associated with numerous services than clinical studies. Code availability was inversely correlated with transparency ($R=-0.36$), as expected (open source was 1 of the assessment conditions).

Figure 1. Overall consensus scores obtained by all studies reviewed. (A) Average consensus scores for all studies reviewed (error bars=SE). (B) Stacked bar graph showing the distribution of scores for each subcomponent question. Ext: external; h/care: health care; int: internal.

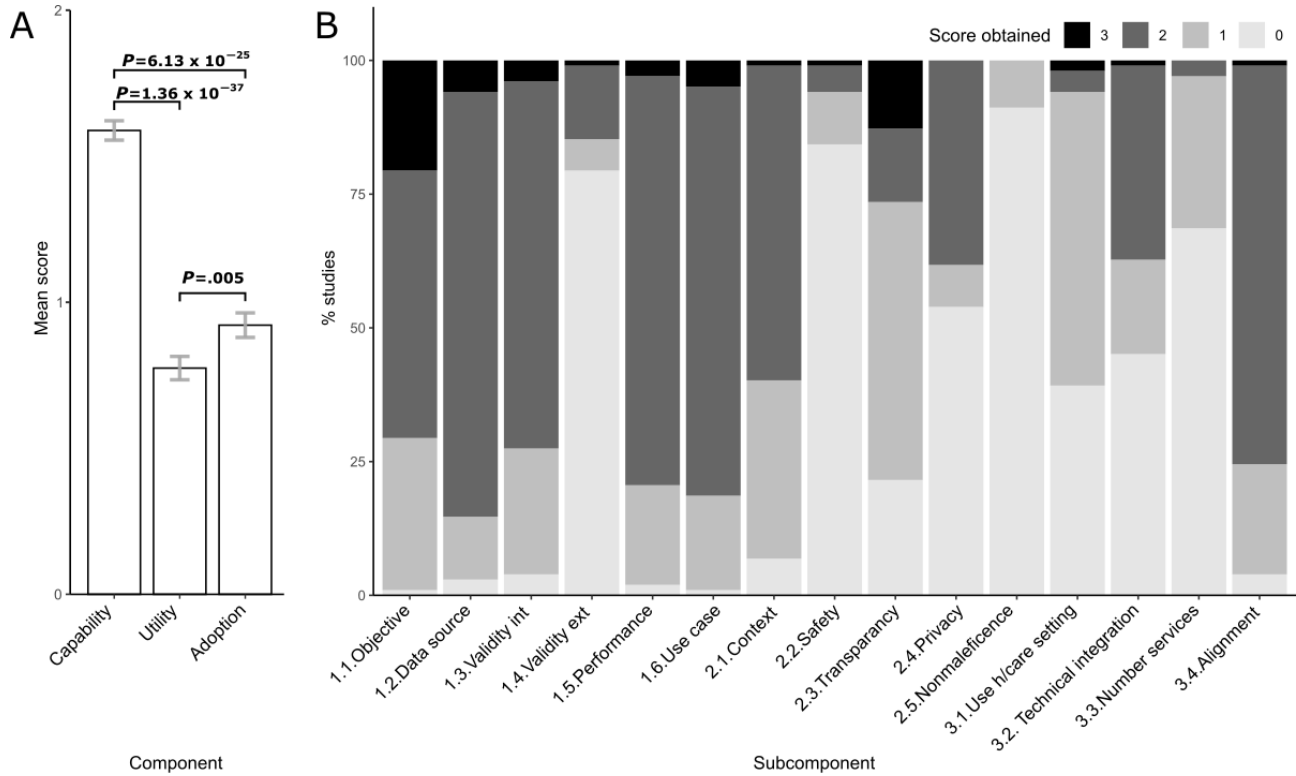
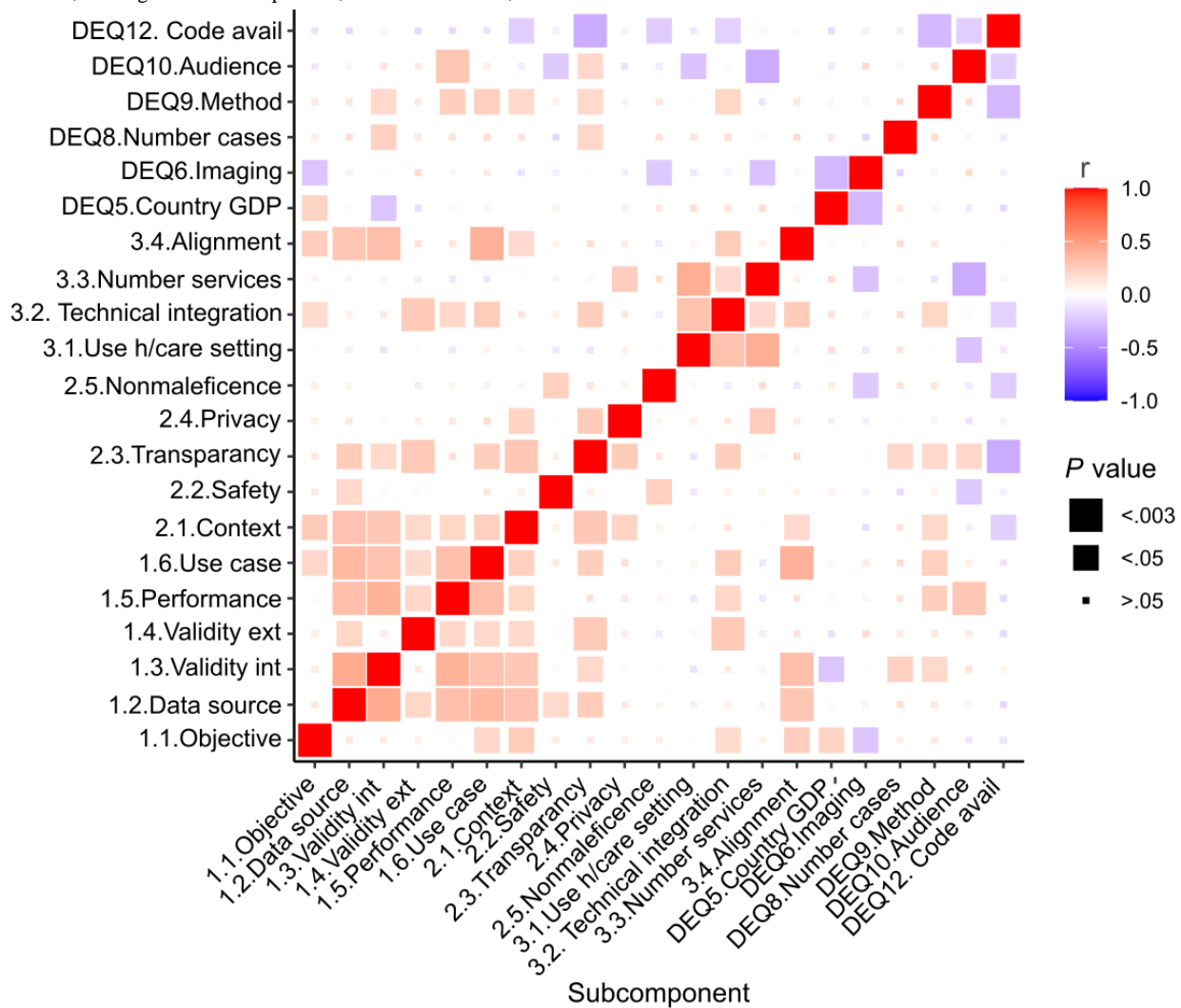


Figure 2. Correlation heatmap showing the strength of correlation between all subcomponents and select data extraction questions. The strength of correlation, as determined by the Fisher exact test, is shown in color, with the size of squares representing the level of significance. Avail: availability; ext: external; GDP: gross domestic product; h/care: health care; int: internal.



AI Study Characteristics

The associations between the AI algorithms used in the studies and TEHAI scores are shown in Figure 3. Deep learning (including a convolutional neural network, or CNN for short) was the most frequent machine learning model (54/102, 52.9%, studies), followed by classic methods (14/102, 13.7%, studies, comprising primarily linear and logistic regression models) and standard machine learning (9/102, 8.8%, studies, comprising primarily random forest [RF] and support vector machine [SVM] algorithms); see Figure 3A. In 20.4% (n=20) of the studies, multiple types of algorithms were used. At the component level, deep learning and machine learning scored better in capability: mean score 1.69 (SE 0.04) and 1.54 (SE 0.12), respectively. In addition, deep learning was superior in adoption: mean score 0.95 (SE 0.06); see Figure 3B. This pattern was also evident at the subcomponent level, where classic methods scored the poorest for most questions (mean scores 0.07-1.78, SE 0.07-0.1), with deep learning scoring significantly higher in numerous subcomponents (mean scores 0.05-1.96, SE 0.03-0.12); see Figure 3C. These findings revealed that those using deep learning are more likely to include facets into their design that

are more likely to ensure their work will be integrated into practice.

Figure 4 contains the results of comparisons between clinical and methodologically focused papers. Methodological studies tended to score higher in the capability component (methodological mean score 1.63, SE 0.04; clinical mean score 1.52, SE 0.06), and clinically focused studies tended to score higher in utility (clinical mean score 0.81, SE 0.07; methodological mean score 0.75, SE 0.05) and adoption (clinical mean score 1.03, SE 0.07; methodological mean score 0.87, SE 0.05; see Figure 4A), particularly in the “use in a health care setting” (clinically focused mean score 0.90, SE 0.11; methodologically focused mean score 0.58, SE 0.08; $P=.037$) and “number of services” (clinically focused mean score 0.58, SE 0.09; methodologically focused mean score 0.23, SE 0.06; $P=2.39 \times 10^{-05}$) subcomponents. It is important to note that all papers scored poorly in the “safety and quality” (clinically focused mean score 0.13, SE 0.14; methodologically focused mean score 0.58, SE 0.05) and “nonmaleficence” (clinically focused mean score 0.12, SE 0.06; methodologically focused mean score 0.07, SE 0.03) subcomponents, and despite being more integrated into the health system, clinical papers did not

score significantly higher scores in these subcomponents (Figures 4A and 4B).

Figure 3. Methods used by the various studies to achieve end points. (A) Percentage of studies using specific methods. As the field of potential algorithms is diverse, we created broad categories to make the pie chart readable and to provide an overview of the most prevalent types of algorithms. Classic methods included linear and logistic regression models, and the machine learning category comprised a heterogeneous mix of established nonlinear algorithms, such as a random forest (RF) and a support vector machine (SVM). The deep learning category included mostly CNNs and represented more recent neural network techniques developed for big data. (B) Component scores for the 4 main methods used in the studies. (C) Subcomponent scores for the 4 main methods used in the studies. Bars show average scores, with error bars equal to SE. Bold *P* values indicate $P < .05$. Bonferroni-corrected significance $P = .003$. CNN: convolutional neural network; ext: external; h/care: health care; int: internal.

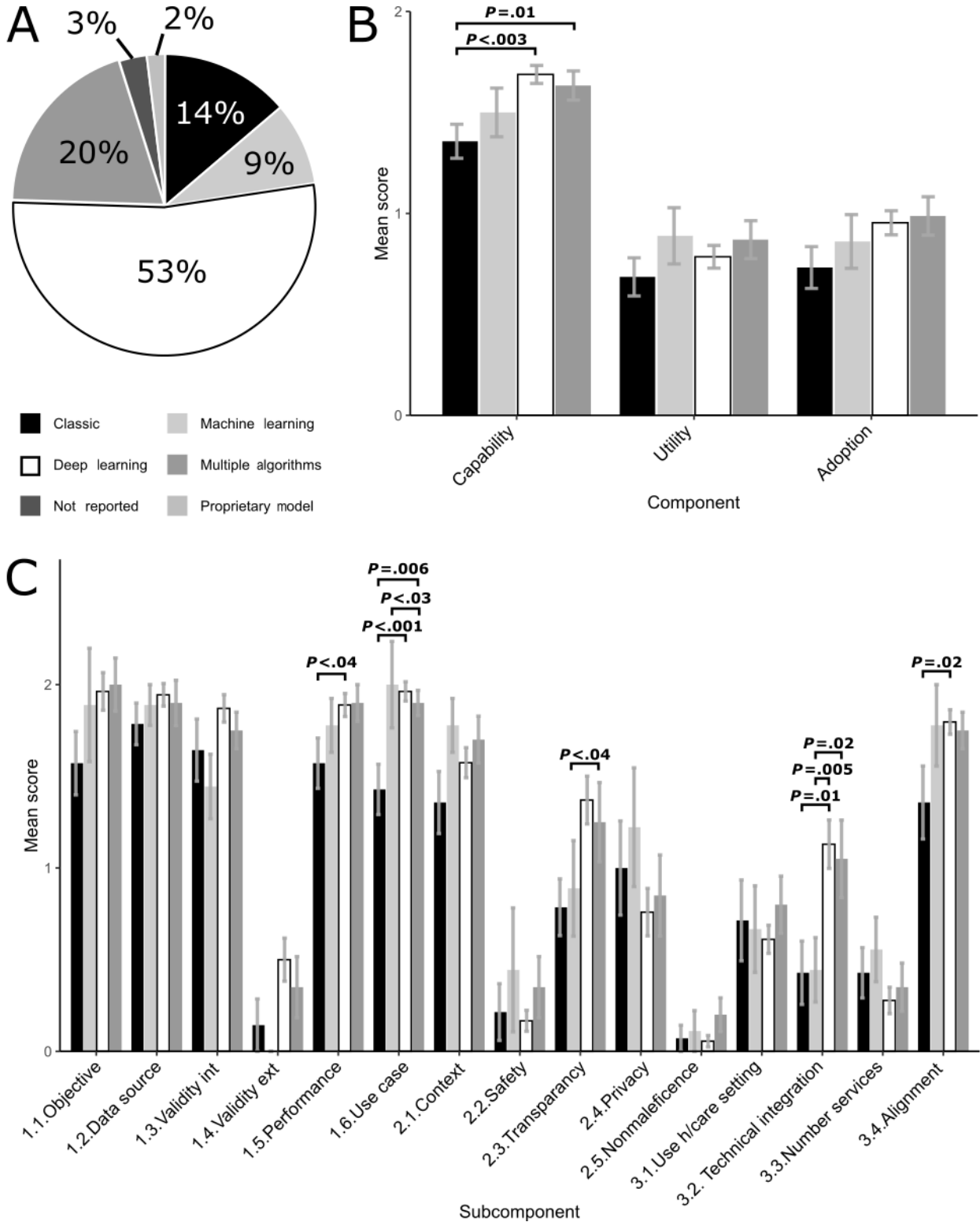
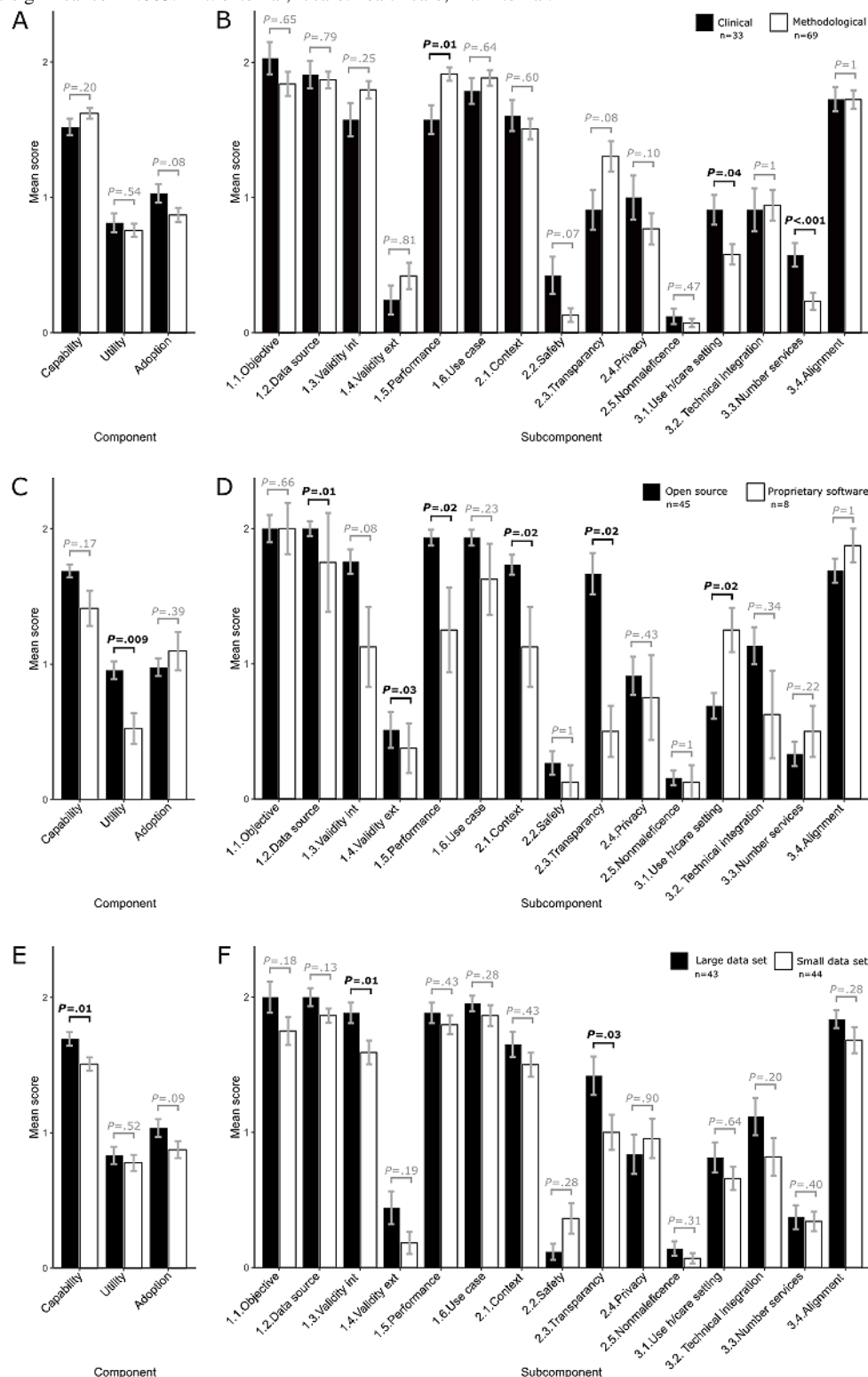


Figure 4. Component and subcomponent scores split into subcategories based on data extraction questions, including (A and B) “intended audience,” (C and D) “type of software,” and (E and F) “size of data set.” Bars show average scores, with error bars equal to SE. Bold *P* values indicate $P < .05$. Bonferroni-corrected significance $P = .003$. Ext: external; h/care: health care; int: internal.



Close to half of the studies used open source software (n=45, 44.1%), with a small portion (n=8, 7.8%) using proprietary software (with the remaining studies being unclear as to the software availability). There was a tendency for proprietary software to perform better at adoption, particularly in the “use in a health care setting” subcomponent (open source software studies mean score 0.69, SE 0.09; proprietary software studies

mean score 1.25, SE 0.16; $P = .02$), while papers with open source software tended to score better in utility, including the “safety and quality” (open source software studies mean score 0.27, SE 0.09; proprietary software studies mean score 0.13, SE 0.13; $P = .99$), “privacy” (open source software studies mean score 0.91, SE 0.14; proprietary software studies mean score 0.75, SE 0.31; $P = .43$), and “nonmaleficence” (open source software

studies mean score 0.15, SE 0.05; proprietary software studies mean score 0.13, SE 0.16; $P=.99$) subcomponents; see [Figures 4C and 4D](#). We also observed a tendency for open source software to score better in the “transparency” subcomponent (open source software studies mean score 1.67, SE 0.15; proprietary software studies mean score 0.5, SE 0.19; $P=.02$), which is compatible with the findings from the correlation analysis ([Figure 2](#)).

Across the studies, the median number of cases was 225 subjects; therefore, we allotted studies with >225 cases to the large-data-set category and those with ≤ 225 cases to the small-data-set category ([Figures 4E and 4F](#)). There was an overall suggestive pattern for the large data set to score higher than the small data set, again with the exception of safety and quality, and both scored poorly in nonmaleficence.

Countries may have differing capacities to integrate new technologies into their health systems, and we hypothesized that it would be detectable via the GDP. We split the studies into low-, middle- and high-income countries based on the classification defined by the World Bank [25]. There were no studies published in the low-income category, with half of the studies originating in middle-income countries and the other half in high-income countries. Interestingly there was no significant difference between components at the multiple testing threshold; however, there was a trend suggesting a difference in the adoption component (high-income study mean score 1.0, SE 0.06; medium-income study mean score 0.83, SE 0.06; $P=.04$; [Multimedia Appendix 4A,B](#)) and a slight tendency toward middle-income countries to score better in the “capability” subcomponent questions, particularly the “objective of the study” (high-income study mean score 2.1, SE 0.09; medium-income study mean score 1.76, SE 0.1; $P=.03$) and “internal validity” (high-income study mean score 1.58, SE 0.08; medium-income study mean score 1.88, SE 0.08; $P=.04$) subcomponents ([Multimedia Appendix 4B](#)).

We found that there were many studies where the authors used AI to analyze images of the lungs (eg, X-rays) of patients with COVID-19 and controls to classify them into categories, ultimately producing algorithms that could accurately identify patients with COVID-19 from images of their lungs. Thus, we classified the studies as being imaging (direct image analysis of X-rays or CT scans) or nonimaging (eg, studies that analyzed blood metabolites), and there was a strong trend for nonimaging studies to score higher than imaging studies, which included the “objective of the study” (imaging study mean score 1.79, SE 0.08; nonimaging study mean score 2.18, SE 0.13; $P=.02$), “safety and quality” (imaging study mean score 0.16, SE 0.05; nonimaging study mean score 0.36, SE 0.14; $P=.015$), “nonmaleficence” (imaging study mean score 0.04, SE 0.02; nonimaging study mean score 0.18, SE 0.07; $P=.05$), and “number of services” (imaging study mean score 0.25, SE 0.06; nonimaging study mean score 0.55, SE 0.11; $P=.02$) subcomponents ([Multimedia Appendix 4C,D](#)).

Discussion

Principal Findings

Considering the emergence of the COVID-19 pandemic and the flurry of AI models that were developed to address various aspects of the pandemic, we conducted a systematic review of these AI models regarding their likely success at translation. We observed a significant trend for studies to attain high scores for technical capability but low scores for the areas essential for clinical translatability. Specific questions regarding external model validation, safety, nonmaleficence, and service adoption received failed scores in most studies. Therefore, we identified notable quality gaps in most AI studies of COVID-19 that are likely to have a negative impact on clinical translation.

There have been many claims made of such AI models, including similar or higher accuracy, sensitivity, or specificity compared to human experts [26-28] and real-time results that have been suggested to lead to improved referral adherence [29], but few independent studies have tested these claims. In fact, it is suggested that although the AI models have potential, they are generally unsuitable for clinical use and, if deployed prematurely, could lead to undesirable outcomes, including stress for both patients and the health system, unnecessary intrusive procedures, and even death due to misdiagnosis [5,7]. Of those studies that examined the utility of COVID-19 AI applications, there has not been a comprehensive evaluation of AI in health care models encompassing assessment of their intrinsic capabilities, external performance, and adoption in health care delivery thus far. It is important for the scientific community and relevant stakeholders to understand how many of these AI models are translational in their value and to what degree. To address this gap, we undertook a comprehensive evaluation of COVID-19 AI models that were developed between December 2019 and December 2020. The framework we chose, TEHAI, is a comprehensive evaluation framework developed by a multidisciplinary international team through a vigorous process of review and consultation and systematically assesses AI models for their translational value [15]. To select COVID-19 studies, we conducted a systematic search, and after screening 3830 studies, we selected 102 studies for evaluation. Based on TEHAI, the studies were assessed for their capability, utility, and adoption aspects and scored using a weighted process.

The scale of the studies we screened (over 3000) and the studies eligible for evaluation (over 900) indicated the level of activity in this area despite the limited time frame selected for the evaluation (2019-2020). The evaluation of the 102 studies, although yielding some interesting findings, also had a few expected results. Notable was that most studies, although doing well in the capability component, did not evaluate highly in the utility and adoption components. The latter components assess the “ethical,” “safety and quality,” and “integration with health care service” aspects of the AI model. However, it is not surprising the AI models scored low in these components, given the expediency required to develop and release these models in a pandemic context. This meant the ethical components were not a priority as one would expect in normal times. It was also

not surprising to find that the CNN was the most popular machine learning model, as most of the selected studies related to medical imaging analysis (69/102 studies were imaging studies compared to 33/102 studies that were not), where the technique is widely understood and beginning to be applied in some clinical settings [6,30].

Although there was a consistent trend for studies with large data sets to score higher than those with small data sets, there was no significant difference in any subcomponent between studies with small versus large data sets. This was a surprising finding and indicates that even when studies have collected more data, they advance no further in the utility or adoption fields, and should the total number of studies analyzed be increased, we would expect the difference between the two data sets to become significant. Regarding imaging versus nonimaging, we observed that nonimaging studies scored higher in some adoption and utility subcomponents. We suspect this was due to the more clinical nature of the nonimaging research teams; thus, the papers focused more on issues important to clinical practice. Although there was a tendency for those studies using proprietary software that we expected to be more mature, the authors had not advanced the findings into practice any more than that of open source, algorithm-based studies. Again, we would expect this difference to become significant if the number of studies scored were to be increased. We also assessed the interpretability of the models as part of the “transparency” subcomponent and found that imaging studies in particular included additional visualization to pinpoint the regions that were driving the classification. Further, the scoring studies in each of the TEHAI components evidenced the need for planning in advance for external validation, safety, and integration in health services to ensure the full translatability of AI models in health care.

Most of the reviewed studies lacked sufficient considerations for adoption into health care practices (the third TEHAI component), which has implications for the business case for AI applications in health care. The cost of deployment and costs from misclassification from both monetary and patient safety/discomfort perspectives can only be assessed if there are pilot data available from actual tests that put new tools into service. Furthermore, critical administrative outcomes, such as workload requirements, should be considered as early as possible. Although we understand that such tests are hard to organize from an academic basis, the TEHAI framework can be used as an incentive to move in this direction.

We note that availability of dedicated data sets and computing resources for training could be a bottleneck for some applications. In this study, we observed multiple instances of transfer learning, which is 1 solution; however, we will revise the capability section of TEHAI to make a more specific consideration for these issues.

Fair access to AI technology should also be part of good design. The TEHAI framework includes this in the “internal validity” subcomponent, where small studies in particular struggled with

representing a sufficient diversity of individuals. From a translational point of view, we also observed shortcomings in the contextualization of AI models. Again, since there was limited evidence on service deployment, most studies scored low on fairness simply due to a lack of data. We also note that deployment in this case may be hindered by the clinical acceptance of the models [11], and we will include this topic in future amendments to the TEHAI framework.

Limitations

Although we undertook a comprehensive evaluation of AI studies unlike previous assessments, our study still has some limitations. First, the period we used to review and select studies was narrow, being just a year. Another limitation is that for practical reasons, we randomly chose a subset of 102 studies for evaluation out of the 968 eligible studies. Despite these limitations, we are confident that the evaluation process we undertook was rigorous, as evidenced by the systematic review of the literature, the detailed assessment of each of the selected studies, and the parallel review and consensus steps.

We recommend caution when generalizing the results from this COVID-19 study to other areas of AI in health care. First, evaluation frameworks that rely on human experts can be sensitive to the selection of the experts (subjectivity). Second, scoring variation may arise from the nature of the clinical problem rather than the AI solution per se; thus, TEHAI results from different fields may not be directly comparable. Third, we intentionally excluded discovery studies aimed at new biology or novel treatments, as those would have been too early in the translation pipeline to have a meaningful evaluation. Fourth, significant heterogeneity of clinical domains may also confound the evaluation results and may prevent comparisons of studies (here, we made an effort to preselect studies that were comparable). Lastly, the TEHAI framework is designed to be widely applicable, which means that stakeholders with specific subjective requirements may need to adapt their interpretations accordingly.

We acknowledge the rapid progress in AI algorithms that may make some of the evaluation aspects obsolete over time; however, we also emphasize that 2 of the 3 TEHAI components are not related to AI itself but to the ways AI interacts with the requirements of clinical practice and health care processes. Therefore, we expect that the translatability observations from this study will have longevity.

Conclusion

AI in health care has a translatability challenge, as evidenced by our evaluation study. By assessing 102 AI studies for their capability, utility, and adoption aspects, we uncovered translational gaps in many of these studies. Our study highlights the need to plan for translational aspects early in the AI development cycle. The evaluation framework we used and the findings from its application will inform developers, researchers, clinicians, authorities, and other stakeholders to develop and deploy more translatable AI models in health care.

Acknowledgments

BK extracted appropriate studies from databases. AEC assigned studies to reviewers, carried out all analysis, and generated figures. All authors were involved in the scoring process. AEC, SR, SA, and V-PM drafted the manuscript. All authors provided feedback and edits for the final manuscript.

Conflicts of Interest

SR holds directorship in Medi-AI. The other authors have no conflicts of interest to declare.

Multimedia Appendix 1

Component and subcomponent scores split into subcategories based on data extraction questions, including (A and B) "country GDP" and (C and D) "imaging/nonimaging"-based study. Bars show average scores, with error bars equal to SE. Bold *P* values indicate $P < .05$. Bonferroni-corrected significance $P = .003$. GDP: gross domestic product.

[[PNG File , 144 KB - ai_v2i1e42313_app1.png](#)]

Multimedia Appendix 2

Search strategies.

[[DOCX File , 15 KB - ai_v2i1e42313_app2.docx](#)]

Multimedia Appendix 3

PRISMA flow diagram.

[[DOCX File , 41 KB - ai_v2i1e42313_app3.docx](#)]

Multimedia Appendix 4

Evaluation and scoring questions.

[[DOCX File , 29 KB - ai_v2i1e42313_app4.docx](#)]

References

1. Desai AN. Artificial intelligence: promise, pitfalls, and perspective. *JAMA* 2020 Jul 23;323(24):2448-2449. [doi: [10.1001/jama.2020.8737](https://doi.org/10.1001/jama.2020.8737)] [Medline: [32492093](https://pubmed.ncbi.nlm.nih.gov/32492093/)]
2. Reddy S, Fox J, Purohit MP. Artificial intelligence-enabled healthcare delivery. *J R Soc Med* 2019 Jan;112(1):22-28 [FREE Full text] [doi: [10.1177/0141076818815510](https://doi.org/10.1177/0141076818815510)] [Medline: [30507284](https://pubmed.ncbi.nlm.nih.gov/30507284/)]
3. Artificial intelligence in health care: benefits and challenges of technologies to augment patient care. United States Government Accountability Office. 2020. URL: <https://www.gao.gov/products/gao-21-7sp> [accessed 2022-07-13]
4. Feng J, Phillips RV, Malenica I, Bishara A, Hubbard AE, Celi LA, et al. Clinical artificial intelligence quality improvement: towards continual monitoring and updating of AI algorithms in healthcare. *NPJ Digit Med* 2022 May 31;5(1):66 [FREE Full text] [doi: [10.1038/s41746-022-00611-y](https://doi.org/10.1038/s41746-022-00611-y)] [Medline: [35641814](https://pubmed.ncbi.nlm.nih.gov/35641814/)]
5. Nsoesie EO. Evaluating artificial intelligence applications in clinical settings. *JAMA Netw Open* 2018 Oct 07;1(5):e182658 [FREE Full text] [doi: [10.1001/jamanetworkopen.2018.2658](https://doi.org/10.1001/jamanetworkopen.2018.2658)] [Medline: [30646173](https://pubmed.ncbi.nlm.nih.gov/30646173/)]
6. van Leeuwen KG, Schalekamp S, Rutten MJCM, van Ginneken B, de Rooij M. Artificial intelligence in radiology: 100 commercially available products and their scientific evidence. *Eur Radiol* 2021 Jul;31(6):3797-3804 [FREE Full text] [doi: [10.1007/s00330-021-07892-z](https://doi.org/10.1007/s00330-021-07892-z)] [Medline: [33856519](https://pubmed.ncbi.nlm.nih.gov/33856519/)]
7. Roberts M, Driggs D, Thorpe M, Gilbey J, Yeung M, Ursprung S, et al. Common pitfalls and recommendations for using machine learning to detect and prognosticate for COVID-19 using chest radiographs and CT scans. *Nat Mach Intell* 2021 Mar 15;3(3):199-217. [doi: [10.1038/s42256-021-00307-0](https://doi.org/10.1038/s42256-021-00307-0)]
8. Seneviratne MG, Shah NH, Chu L. Bridging the implementation gap of machine learning in healthcare. *BMJ Innov* 2019 Dec 20;6(2):45-47. [doi: [10.1136/bmjinnov-2019-000359](https://doi.org/10.1136/bmjinnov-2019-000359)]
9. Kim DW, Jang HY, Kim KW, Shin Y, Park SH. Design characteristics of studies reporting the performance of artificial intelligence algorithms for diagnostic analysis of medical images: results from recently published papers. *Korean J Radiol* 2019 Mar;20(3):405-410 [FREE Full text] [doi: [10.3348/kjr.2019.0025](https://doi.org/10.3348/kjr.2019.0025)] [Medline: [30799571](https://pubmed.ncbi.nlm.nih.gov/30799571/)]
10. Yu AC, Mohajer B, Eng J. External validation of deep learning algorithms for radiologic diagnosis: a systematic review. *Radiol Artif Intell* 2022 May;4(3):e210064 [FREE Full text] [doi: [10.1148/ryai.210064](https://doi.org/10.1148/ryai.210064)] [Medline: [35652114](https://pubmed.ncbi.nlm.nih.gov/35652114/)]
11. Schneider J, Agus M. Reflections on the clinical acceptance of artificial intelligence. In: Househ M, Borycki E, Kushniruk A, editors. *Multiple Perspectives on Artificial Intelligence in Healthcare: Opportunities and Challenges*. Cham: Springer International Publishing; 2021:103-114.
12. Reddy S, Allan S, Coghlan S, Cooper P. A governance model for the application of AI in health care. *J Am Med Inform Assoc* 2020 Mar 01;27(3):491-497 [FREE Full text] [doi: [10.1093/jamia/ocz192](https://doi.org/10.1093/jamia/ocz192)] [Medline: [31682262](https://pubmed.ncbi.nlm.nih.gov/31682262/)]

13. Mhasawade V, Zhao Y, Chunara R. Machine learning and algorithmic fairness in public and population health. *Nat Mach Intell* 2021 Jul 29;3(8):659-666. [doi: [10.1038/s42256-021-00373-4](https://doi.org/10.1038/s42256-021-00373-4)]
14. AlHasan A. Bias in medical artificial intelligence. *Bull R Coll Surg Engl* 2021 Sep;103(6):302-305. [doi: [10.1308/rcsbull.2021.111](https://doi.org/10.1308/rcsbull.2021.111)]
15. Reddy S, Rogers W, Makinen V, Coiera E, Brown P, Wenzel M, et al. Evaluation framework to guide implementation of AI systems into healthcare settings. *BMJ Health Care Inform* 2021 Oct;28(1):e100444 [FREE Full text] [doi: [10.1136/bmjhci-2021-100444](https://doi.org/10.1136/bmjhci-2021-100444)] [Medline: [34642177](https://pubmed.ncbi.nlm.nih.gov/34642177/)]
16. Kim W, Jang Y, Yang J, Chung J. Spatial activation of TORC1 is regulated by Hedgehog and E2F1 signaling in the *Drosophila* eye. *Dev Cell* 2017 Aug 21;42(4):363-375.e4 [FREE Full text] [doi: [10.1016/j.devcel.2017.07.020](https://doi.org/10.1016/j.devcel.2017.07.020)] [Medline: [28829944](https://pubmed.ncbi.nlm.nih.gov/28829944/)]
17. Saygılı A. A new approach for computer-aided detection of coronavirus (COVID-19) from CT and X-ray images using machine learning methods. *Appl Soft Comput* 2021 Jul;105:107323 [FREE Full text] [doi: [10.1016/j.asoc.2021.107323](https://doi.org/10.1016/j.asoc.2021.107323)] [Medline: [33746657](https://pubmed.ncbi.nlm.nih.gov/33746657/)]
18. Roimi M, Gutman R, Somer J, Ben Arie A, Calman I, Bar-Lavie Y, et al. Development and validation of a machine learning model predicting illness trajectory and hospital utilization of COVID-19 patients: a nationwide study. *J Am Med Inform Assoc* 2021 Jul 12;28(6):1188-1196 [FREE Full text] [doi: [10.1093/jamia/ocab005](https://doi.org/10.1093/jamia/ocab005)] [Medline: [33479727](https://pubmed.ncbi.nlm.nih.gov/33479727/)]
19. Jin C, Chen W, Cao Y, Xu Z, Tan Z, Zhang X, et al. Development and evaluation of an artificial intelligence system for COVID-19 diagnosis. *Nat Commun* 2020 Oct 09;11(1):5088 [FREE Full text] [doi: [10.1038/s41467-020-18685-1](https://doi.org/10.1038/s41467-020-18685-1)] [Medline: [33037212](https://pubmed.ncbi.nlm.nih.gov/33037212/)]
20. Reddy S, Bhaskar R, Padmanabhan S, Verspoor K, Mamillapalli C, Lahoti R, et al. Use and validation of text mining and cluster algorithms to derive insights from corona virus disease-2019 (COVID-19) medical literature. *Comput Methods Programs Biomed Update* 2021;1:100010 [FREE Full text] [doi: [10.1016/j.cmpbup.2021.100010](https://doi.org/10.1016/j.cmpbup.2021.100010)] [Medline: [34337589](https://pubmed.ncbi.nlm.nih.gov/34337589/)]
21. Guo Y, Zhang Y, Lyu T, Prosperi M, Wang F, Xu H, et al. The application of artificial intelligence and data integration in COVID-19 studies: a scoping review. *J Am Med Inform Assoc* 2021 Aug 13;28(9):2050-2067 [FREE Full text] [doi: [10.1093/jamia/ocab098](https://doi.org/10.1093/jamia/ocab098)] [Medline: [34151987](https://pubmed.ncbi.nlm.nih.gov/34151987/)]
22. Covidence systematic review software. Veritas Health Innovation. URL: <https://www.covidence.org> [accessed 2023-06-01]
23. Cohen J. A coefficient of agreement for nominal scales. *Educ Psychol Meas* 2016 Jul 02;20(1):37-46. [doi: [10.1177/001316446002000104](https://doi.org/10.1177/001316446002000104)]
24. Sim J, Wright CC. The kappa statistic in reliability studies: use, interpretation, and sample size requirements. *Phys Ther* 2005 Mar;85(3):257-268. [Medline: [15733050](https://pubmed.ncbi.nlm.nih.gov/15733050/)]
25. The world by income and region. The World Bank. 2022. URL: <https://datatopics.worldbank.org/world-development-indicators/the-world-by-income-and-region.html> [accessed 2022-07-13]
26. Ruamviboonsuk P, Tiwari R, Sayres R, Nganthavee V, Hemarat K, Kongprayoon A, et al. Real-time diabetic retinopathy screening by deep learning in a multisite national screening programme: a prospective interventional cohort study. *Lancet Digit Health* 2022 May;4(4):e235-e244 [FREE Full text] [doi: [10.1016/S2589-7500\(22\)00017-6](https://doi.org/10.1016/S2589-7500(22)00017-6)] [Medline: [35272972](https://pubmed.ncbi.nlm.nih.gov/35272972/)]
27. Cen LP, Ji J, Lin J, Ju S, Lin H, Li T, et al. Automatic detection of 39 fundus diseases and conditions in retinal photographs using deep neural networks. *Nat Commun* 2021 Aug 10;12(1):4828 [FREE Full text] [doi: [10.1038/s41467-021-25138-w](https://doi.org/10.1038/s41467-021-25138-w)] [Medline: [34376678](https://pubmed.ncbi.nlm.nih.gov/34376678/)]
28. Deperlioglu O, Kose U, Gupta D, Khanna A, Sangaiyah AK. Diagnosis of heart diseases by a secure Internet of Health Things system based on autoencoder deep neural network. *Comput Commun* 2020 Oct 01;162:31-50 [FREE Full text] [doi: [10.1016/j.comcom.2020.08.011](https://doi.org/10.1016/j.comcom.2020.08.011)] [Medline: [32843778](https://pubmed.ncbi.nlm.nih.gov/32843778/)]
29. Liu J, Gibson E, Ramchal S, Shankar V, Piggott K, Sychev Y, et al. Diabetic retinopathy screening with automated retinal image analysis in a primary care setting improves adherence to ophthalmic care. *Ophthalmol Retina* 2021 Jan;5(1):71-77 [FREE Full text] [doi: [10.1016/j.oret.2020.06.016](https://doi.org/10.1016/j.oret.2020.06.016)] [Medline: [32562885](https://pubmed.ncbi.nlm.nih.gov/32562885/)]
30. Omoumi P, Ducarouge A, Tournier A, Harvey H, Kahn CE, Louvet-de Verchère F, et al. To buy or not to buy-evaluating commercial AI solutions in radiology (the ECLAIR guidelines). *Eur Radiol* 2021 Jul;31(6):3786-3796 [FREE Full text] [doi: [10.1007/s00330-020-07684-x](https://doi.org/10.1007/s00330-020-07684-x)] [Medline: [33666696](https://pubmed.ncbi.nlm.nih.gov/33666696/)]

Abbreviations

- AI:** artificial intelligence
- CNN:** convolutional neural network
- GDP:** gross domestic product
- MeSH:** Medical Subject Headings
- NIH:** National Institutes of Health
- TEHAI:** Translational Evaluation of Healthcare AI

Edited by K El Emam, B Malin; submitted 31.08.22; peer-reviewed by W Klement, S Lin; comments to author 07.11.22; revised version received 23.11.22; accepted 22.03.23; published 06.07.23.

Please cite as:

*Casey AE, Ansari S, Nakisa B, Kelly B, Brown P, Cooper P, Muhammad I, Livingstone S, Reddy S, Makinen VP
Application of a Comprehensive Evaluation Framework to COVID-19 Studies: Systematic Review of Translational Aspects of Artificial Intelligence in Health Care*

JMIR AI 2023;2:e42313

URL: <https://ai.jmir.org/2023/1/e42313>

doi: [10.2196/42313](https://doi.org/10.2196/42313)

PMID: [37457747](https://pubmed.ncbi.nlm.nih.gov/37457747/)

©Aaron Edward Casey, Saba Ansari, Bahareh Nakisa, Blair Kelly, Pieta Brown, Paul Cooper, Imran Muhammad, Steven Livingstone, Sandeep Reddy, Ville-Petteri Makinen. Originally published in JMIR AI (<https://ai.jmir.org>), 06.07.2023. This is an open-access article distributed under the terms of the Creative Commons Attribution License (<https://creativecommons.org/licenses/by/4.0/>), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR AI, is properly cited. The complete bibliographic information, a link to the original publication on <https://www.ai.jmir.org/>, as well as this copyright and license information must be included.

Review

Predicting Adherence to Behavior Change Support Systems Using Machine Learning: Systematic Review

Akon Obu Ekpezu¹, MPhil; Isaac Wiafe², PhD; Harri Oinas-Kukkonen¹, PhD

¹Oulu Advanced Research on Service and Information Systems, Department of Information Processing Science, University of Oulu, Oulu, Finland

²Department of Computer Science, University of Ghana, Accra, Ghana

Corresponding Author:

Akon Obu Ekpezu, MPhil

Oulu Advanced Research on Service and Information Systems

Department of Information Processing Science

University of Oulu

Pentti Kaiteran Katu 1 Linnanmaa

Oulu, 90570

Finland

Phone: 358 468860704

Email: akon.ekpezu@oulu.fi

Abstract

Background: There is a dearth of knowledge on reliable adherence prediction measures in behavior change support systems (BCSSs). Existing reviews have predominately focused on self-reporting measures of adherence. These measures are susceptible to overestimation or underestimation of adherence behavior.

Objective: This systematic review seeks to identify and summarize trends in the use of machine learning approaches to predict adherence to BCSSs.

Methods: Systematic literature searches were conducted in the Scopus and PubMed electronic databases between January 2011 and August 2022. The initial search retrieved 2182 journal papers, but only 11 of these papers were eligible for this review.

Results: A total of 4 categories of adherence problems in BCSSs were identified: adherence to digital cognitive and behavioral interventions, medication adherence, physical activity adherence, and diet adherence. The use of machine learning techniques for real-time adherence prediction in BCSSs is gaining research attention. A total of 13 unique supervised learning techniques were identified and the majority of them were traditional machine learning techniques (eg, support vector machine). Long short-term memory, multilayer perception, and ensemble learning are currently the only advanced learning techniques. Despite the heterogeneity in the feature selection approaches, most prediction models achieved good classification accuracies. This indicates that the features or predictors used were a good representation of the adherence problem.

Conclusions: Using machine learning algorithms to predict the adherence behavior of a BCSS user can facilitate the reinforcement of adherence behavior. This can be achieved by developing intelligent BCSSs that can provide users with more personalized, tailored, and timely suggestions.

(JMIR AI 2023;2:e46779) doi:[10.2196/46779](https://doi.org/10.2196/46779)

KEYWORDS

adherence; compliance; behavior change support systems; persuasive systems; persuasive technology; machine learning

Introduction

Behavior change support systems (BCSSs) have been effective in improving health and healthier lifestyles. These are persuasive systems that have been designed to change behavior without force or deception [1]. However, the effectiveness of these systems is generally hindered by nonadherence [2-4]. Nonadherence to recommended regimes in BCSSs has the

potential to diminish their long-term benefits [5]. It is associated with the increased prevalence of diseases such as hypertension, diabetes, obesity, dementia, bipolar disorder, and heart failure [2,4,6-8], as well as the increased cost of health care. Yet, there are no standardized factors that can reliably predict adherence [9,10]. Direct adherence monitoring approaches are expensive, burdensome to care providers, and susceptible to distortion by patients, while indirect monitoring approaches such as pill count,

patient questionnaires, electronic medication monitors, or electronic reporting of daily physical activity are susceptible to misinterpretations and overestimation of adherence [11,12]. To implement effective BCSSs and ensure positive behavior change outcomes that can be attributed to the recommended interventions, an accurate assessment of adherence behaviors and their predictors has become imperative. This will guide researchers and health care providers in identifying nonadherent individuals as well as provide measures that will re-engage and help them to adhere [13]. Additionally, an early prediction of user dropout or relapse during interventions may suggest measures that can be used to improve adherence [14].

Existing systematic reviews [2,7,15-19] have sought to examine predictors or determinants of adherence to several BCSSs. They predominately report that there is a lack of consistency regarding reports of adherence, key variables mediating adherence, and reliable measures of adherence. However, findings from these reviews were based on studies that relied solely on self-reported measures of adherence using pharmacological claims and validated questionnaires from behavior change and health psychology theories. Hence, abounding issues of over- and underreporting may limit the validity of the findings.

This review enhances existing knowledge by focusing on predictors of adherence to BCSSs using machine learning techniques. Machine learning techniques have enabled a proficient means of classifying, detecting, and predicting complex phenomena including human behavior. It has also attracted considerable research interest in the development of BCSSs [20-22]. Nonetheless, literature on the use of machine learning techniques as adherence prediction methods in BCSSs is limited [13,23]. Although Bohlmann et al [23] provided literature summaries on machine learning techniques for predicting adherence, they focused on medication adherence only and considered both digital and nondigital interventions. In contrast to previous reviews, this systematic review focuses on the use of machine learning approaches to predict all kinds of adherence problems in BCSSs. In addition, it focuses only on primary studies that used objectively collected data or data generated by the BCSS. Accordingly, this review seeks to answer the following question: What are the existing trends in the use of machine learning techniques to predict adherence to BCSSs? Specifically, this study answers 4 main review questions, as shown in Table 1.

Table 1. Review questions (RQs) and their motivations.

RQ	Question	Motivation
RQ1	What are the targeted adherence problems and their related definitions?	Research on adherence has predominately focused on adherence to medication and pharmacological treatments. However, adherence covers a wider range of health behaviors than medication adherence [9]. This RQ sought to identify other target adherence problems in BCSSs ^a .
RQ2	What are the characteristics of the BCSS including persuasive system features?	Considering the variabilities in adherence problems and BCSSs, this RQ aimed to provide summaries on the characteristics of the BCSS and the persuasive system features that have been used to improve adherence.
RQ3	What are the adopted machine learning approaches in predicting adherence to BCSSs?	This RQ sought to identify the nature of the raw data and predominately used machine learning techniques, feature selection techniques, and performance metrics.
RQ4	What are the limitations or barriers to adherence?	Though various barriers to adherence have been identified in the literature, this RQ sought to identify only those barriers that limit individuals from adhering to the request of the BCSS.

^aBCSS: behavior change support system.

Methods

Literature Search

Following the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) approach, a search on Scopus and PubMed electronic databases was conducted. This search aimed to identify peer-reviewed English conference and journal papers published between January 2011 and August 2022. Scopus indexes a larger number of peer-reviewed scientific journals than the Web of Science and offers results of more consistent accuracy than Google Scholar, while PubMed remains a leading database in biomedical research [24]. Including papers published within the past decade will reveal recent evidence-based research trends [25]. Using the logical OR/AND operators, the search phrases were a combination of keywords related to prediction, adherence, health behavior change interventions, and machine learning (See Multimedia Appendix 1 for the search phrases). Considering the plethora of approaches

to investigate adherence, the search for eligible studies was not limited to a specific study design.

Only empirical studies that described the development and testing of machine learning models for BCSS adherence prediction were considered. Studies that used only self-reported data, were not reported in English, or did not focus on human participants were ignored.

Study Selection

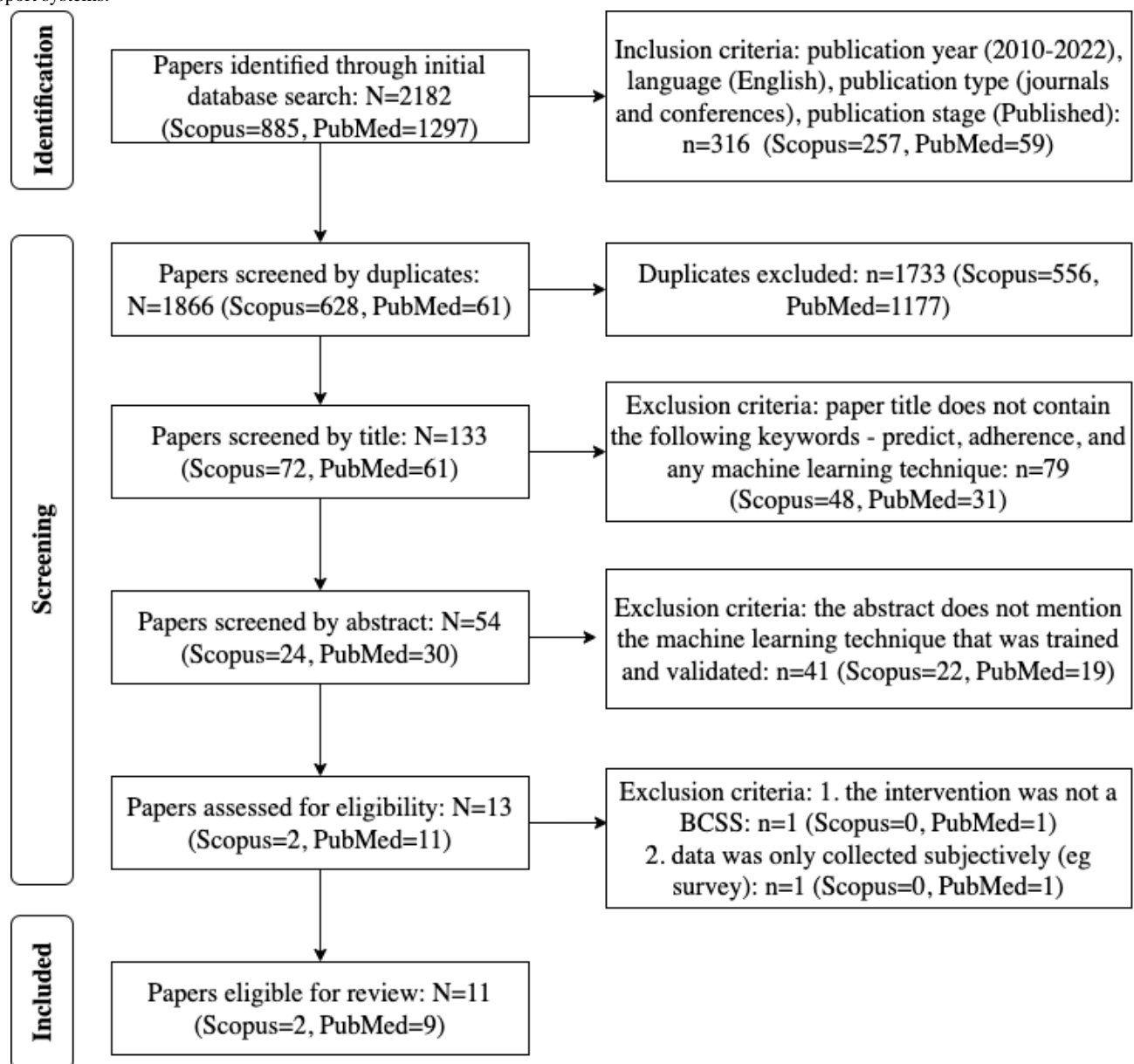
During the initial search of the databases, 2182 papers were retrieved. The results were refined by year, document type, publication stage, source type, and language, resulting in 1866 papers. The exported papers were screened for uniqueness and for titles containing keywords such as adherence, prediction, and any machine learning technique. Of these, 1812 papers were excluded. The remaining 54 papers were screened by abstracts and full texts. Papers were excluded by abstract if the machine learning technique(s) were not mentioned. Furthermore, papers were excluded by full text if the intervention was not

characterized by a BCSS (any form of information system that has been developed to change human behavior voluntarily). Finally, 11 journal papers were considered eligible for this systematic review and thus downloaded for methodological quality assessment.

Figure 1 shows the PRISMA flow diagram of study identification and selection. Since the study selection was not limited to a specific study design, the Mixed Methods Appraisal

Tool by [26] was used to assess the methodological quality of the downloaded papers (see Multimedia Appendix 2 [13,14,26-35]). Accordingly, 11 studies were identified to be of high methodological quality. Pertinent information on the study characteristics and machine learning approaches was extracted using a data extraction form in Microsoft Excel created by the authors. Multimedia Appendix 3 [13,14,27-35] presents a list of included studies.

Figure 1. PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) flow diagram for study selection. BCSS: behavior change support systems.



Results

All included studies (N=11) were primarily aimed at the development and use of machine learning techniques for the prediction of adherence [27-32], nonadherence, dropout, or relapse [13,14,33-35]. The ensuing sections will elaborate on findings related to the review questions (RQs).

Targeted Adherence Problems and Their Related Definitions (RQ1)

Overview

This review identified 7 health behaviors that BCSSs target: medication adherence [29,31,34], use of health care systems [13,28], physical activity [14,27], diet [33], illicit drug use [30], depression and anxiety [32], and insomnia [35]. Based on the characteristics of these health behaviors, they were grouped into

4 categories of adherence problems, described in the following sections.

Adherence to Digital Cognitive or Behavioral Interventions (n=5)

This category includes health behaviors such as health care system use, illicit drug use, depression, anxiety, and insomnia. This adherence problem focused on predicting adherence or nonadherence to internet-based cognitive behavioral therapy [32,35], and remote health monitoring systems [13,28,30] using machine learning models. Adherence to digital cognitive or behavioral interventions refers to the successful completion of all recommended tasks and achievement of initial set goals. While nonadherence refers to ignoring or not completing the recommended task consecutively after signing up to use the BCSS.

Medication Adherence (n=3)

This adherence problem is linked to medication adherence behavior. Although extant literature posit that medication nonadherence is the most common form of adherence problem, this review identified 3 studies that addressed this problem within BCSSs. It compromises the effectiveness of treatment outcomes in about 85% of patients with chronic and acute medical conditions globally [11,36]. Studies in this category applied machine learning models to remote real-time measurements of medication dosing [29,31,34]. Though these studies had different thresholds for defining medication adherence, it generally referred to a patient's behavior or commitment to taking the medications as prescribed by a physician with an average adherence rate of 80% and above.

Physical Activity Adherence (n=2)

Adhering to physical activity routines has the potential to reduce the risk of chronic diseases irrespective of age or other sociodemographic factors. Whereas some individuals find it difficult to regularly engage in or continue a physical activity routine [14], others discontinue when they have achieved a health or body goal [27]. These studies were observed to have

varying definitions of adherence. For instance, Zhou et al [14] considered an increase in the number of steps over time, while Bastidas et al [27] considered the users' app use patterns. Thus, physical activity adherence may be defined as either a consistent increase in physical activity levels compared to an individual's baseline activity levels or an individual's responsiveness to prompts from the app. These 2 definitions describe behavior compliance and program compliance, respectively [37].

Diet Adherence (n=1)

This was observed in only 1 study [33]. Dietary relapse in a weight loss intervention was predicted. Dietary relapse refers to any instance in which a person exceeded a specified meal or snack point threshold (per meal).

Characteristics of the BCSSs and Persuasive System Features (RQ2)

The BCSSs included mobile apps [13,14,27,28,33,34], web-based apps [30,32,33,35], and sensor-based systems plus mobile apps [29,31]. They were targeted at different groups of people, namely physically inactive women, illicit drug users, obese or overweight people, and patients with a wide range of chronic diseases, such as heart failure, myocardial infarction-anxiety, depression (MI-ANXDEP), insomnia, and Parkinson disease. [Multimedia Appendix 4](#) [14,15,29-44] describes other study-specific characteristics.

The BCSSs leveraged some behavior change techniques and persuasive systems features to improve user adherence. These features were extracted and evaluated using the persuasive systems design (PSD) model [45]. The PSD model has been validated in several studies [37] and is predominately used in the design and evaluation of BCSSs [46]. The model consists of 28 system features that make up 4 categories of persuasive principles (namely primary task support, dialogue support, credibility support, and social support). [Table 2](#) displays the frequency of the PSD features represented in the BCSS. All the studies had primary task support features, 8 studies had dialogue support features, 5 studies had credibility support features, and only 1 study had social support features.

Table 2. Persuasive features identified in the behavior change support systems (BCSSs). Check marks indicate that the feature was identified.

Persuasive features	Studies predicting adherence					Studies predicting nonadherence					Total	
	[30]	[28]	[32]	[29]	[31]	[27]	[13]	[35]	[33]	[34]		[14]
Primary task support												
Personalization	✓		✓	✓	✓	✓	✓	✓		✓	✓	8/11
Self-monitoring		✓				✓	✓				✓	4/11
Reduction				✓		✓			✓	✓		4/11
Rehearsal					✓							1/11
Tailoring	✓											1/11
Dialogue support												
Reminders		✓		✓				✓	✓	✓		5/11
Suggestions							✓				✓	2/11
Praise	✓										✓	2/11
Rewards									✓			1/11
Similarity	✓											1/11
Liking								✓				1/11
Credibility support												
Real world feel	✓					✓		✓		✓	✓	5/11
Expertise	✓					✓		✓				3/11
Verifiability	✓							✓		✓		3/11
Third party endorsement	✓											1/11
Social support												
Social facilitation	✓						✓					2/11

Primary task support simplifies and motivates users to perform recommended tasks (eg, exercise). A total of 5 features of the primary task support principle were used: personalization, self-monitoring, reduction, rehearsals, and tailoring. Personalization was the most used feature in this category. It delivers personalized content to the users. For example, artificial intelligence generated workouts according to an individual's characteristics and preferences. Self-monitoring enables app users to view and track their activity levels and health status in real-time (eg, the app enables users to monitor, visualize, and track activity levels and calories burned in real time). Reduction breaks down tasks such as daily point goals into specific meal or snack targets. The least used features in this category were rehearsal (practicing the target behavior, eg, gait movements) and tailoring (eg, the app provided content that was distinct to users of specific age groups and health goals, such as alcohol or smoking cessation, weight loss, or mental health).

Dialogue support provides a means to help users to achieve their goals via human-computer interactions. The dialogue support features included reminders (eg, medication prompts), suggestions (eg, the app advises users based on their input to the app), and praise (automated feedback on the completion of a task). The least used features included rewards (eg, point-based incentives), similarity (eg, therapy resembling traditional cognitive behavioral therapy), and liking (eg, user-friendly and appealing design).

System credibility support provides a means for users to trust the system. Features identified in this category included expertise (app provided theory-based information and were designed to improve engagement, effectiveness, and security), verifiability (app provided links to related sites), third-party endorsement (from the National Institute of Health), and real-world feel. The real-world feel feature was implemented as "Contact Us" (a means to communicate with the developers of the app) and "About Us" (providing information about the developers of the app).

Social support provides a means of supporting users via social influence. However, it was the least used principle. Social facilitation was the only identified social support feature and it was implemented by allowing user participation in online app forums. Perhaps, the minimal use of social support features may be attributed to the negative sentiments associated with some of its features [38].

It is important to note that the effectiveness of these persuasive system features in improving adherence was not explicitly evaluated in the included studies. However, these features may have directly or indirectly improved adherence rates. Some of the studies used behavior change techniques such as goal-setting, web-based human coaching, face-to-face counseling sessions, and feedback from psychologists and expert program providers to improve adherence behavior. However, this review could not

identify behavioral theories or models upon which these techniques were based.

Machine Learning Approaches (RQ3)

Developing machine learning models for predicting user adherence or nonadherence was the general aim of all the included studies. This process conventionally consists of 4 main stages: data collection, feature selection, model training, and model validation.

Data Collection

Apart from the gait-related data, which were collected in a controlled environment (laboratory), the data used to train or test the machine learning models in the majority of the studies were objectively collected by the health app while the study participants were performing the behavior of interest in an uncontrolled environment. For example, data such as log data, training behavior, walking steps, and 3D movement scans were automatically extracted in a contactless manner without any form of self-report from the users.

Studies on medication adherence were observed to use images and videos to capture participants' medication adherence behavior. The apps used technologies such as computer vision [34], internet-connected smart sharp bin [29], and flight sensors [31] for the real-time monitoring of self-administered injections and medication ingestion. Time-stamped data of injection needles discarded into the smart sharp bin, time-stamped skeletal joint data, and images of the participants taking the medication were retrieved, validated, and then used to generate the data set for training or testing the model. Similarly, studies on physical activity adherence [14,27] used continuously collected time-dependent physical activity data from app users to develop machine learning models.

Studies on adherence to digital cognitive or behavioral interventions [13,28,30,32,35] and dietary adherence [33] used a combination of objectively collected data and participants' responses to questions on self-assessment delivered by the health

app. Goldstein et al [33] used a BCSS that asked predefined questions related to triggers of dietary relapse for the analysis. Considering that objectively collecting trigger-related data or self-assessment data on health symptoms from a mobile app may currently be challenging as triggers (such as food cravings and hunger) are physiologically motivated, future studies on BCSSs that seek to extract trigger-related data may consider using physiological sensor data. This is a noninvasive approach to detecting hunger and cravings using wearable body sensors [39]. Such sensor data may also be integrated into the health app to enable self-monitoring.

Feature Selection or Engineering

Among the 11 studies, 5 performed feature selection, 5 performed feature engineering, and 1 adopted features based on existing literature (see [Multimedia Appendix 5](#) [14,15,29-37]). Feature selection and feature engineering were both aimed at enhancing model performance by eliminating irrelevant features and generating new features from raw data respectively. Due to the complexities of combining 2 or more machine learning techniques (ie, ensemble learning), some studies [28] applied more than 1 feature selection method. However, there were no differences in the selected features.

The flat features algorithm (including filter, wrapper, and embedded methods) were the predominately used feature selection method. This algorithm assumes all features to be independent [40]. Interestingly, each study had its own set of unique predictors irrespective of the category of adherence problem. [Multimedia Appendix 5](#) [14,15,29-37] highlights the various feature selection approaches.

Model Training

The learning problem was a binary classification. Thus, there were 2 class labels (outcomes), namely adherence/nonadherence, adherers/nonadherers, relapse/nonrelapse, and dropout/nondropout. An overview of the adopted techniques and the outcomes of the best-performing techniques is provided in [Table 3](#).

Table 3. Identified machine learning and model validation techniques.

Ref	Machine learning techniques	Evaluation metrics	Predicted outcome
[30]	Logistic regression and random forest ^a	AUROC ^{a,b} , specificity, sensitivity, PPV ^c , NPV ^d , and confusion matrix	Successful or early dropout
[13]	Logistic regression, random forest ^a , and decision trees	Accuracy, precision ^a , and AUROC	Dropout or nondropout
[28]	Decision tree, MLP ^e , and KNN ^{a,f}	Precision, sensitivity, F_1 -score, TPR ^g , FPR ^h , and AUROC ^a	Adherers or nonadherers
[32]	Random forest ^a	Accuracy	Adherence or nonadherence
[35]	Logistic regression, SVM ⁱ , and decision trees (boosted) ^a	AUROC ^a , TPR, FPR, and PRAUC ^j	Dropout or nondropout
[33]	Ensemble methods ^a	Accuracy, sensitivity ^a , specificity, and AUROC ^a	Relapse or not
[29]	XGB ^k , extra trees, random forest, MLP, gradient tree boosting, recurrent neural network, and LSTM ^{a,l}	Accuracy, specificity ^a , sensitivity, precision, F_1 -score, and AUROC	Adherence or nonadherence
[34]	XGB ^a	Accuracy, precision ^a , sensitivity, AUROC, TPR, and FPR	Adherence or nonadherence
[31]	Decision trees ^a , KNN, naive Bayes, SVM, and random forest	Confusion matrix	Adherence or nonadherence
[27]	LSTM ^a and SVM	Accuracy, sensitivity, F_1 -score, and confusion matrix	Adherent or nonadherent
[14]	Logistic regression ^a and SVM	Accuracy, sensitivity, specificity, and AUROC	Relapse or not

^aBest performing machine learning technique or most relevant metric for the outcome prediction.

^bAUROC: area under the receiver operating characteristic curve.

^cPPV: positive predictive value.

^dNPV: negative predictive value.

^eMLP: multilayer perceptron.

^fKNN: k-nearest neighbor.

^gTPR: true positive rate.

^hFPR: false positive rate.

ⁱSVM: support vector machine.

^jPRAUC: precision-recall curve.

^kXGB: extreme gradient boost.

^lLSTM: long short-term memory.

A total of 13 supervised machine-learning techniques were used across the included studies. Logistic regression, support vector machines, and random forest were the most used techniques cutting across all 4 categories of the adherence problems. The machine learning techniques mapped to specific adherence problems included support vector machines for physical activity adherence; extreme gradient boosting, extra trees, recurrent neural network, naive Bayes, and gradient tree boosting for medication adherence; and ensemble methods for dietary adherence. Random forest was observed to be the predominant best-performing model in studies on adherence to digital cognitive or behavioral interventions, while long short-term memory (LSTM) was a common best-performing model between medication adherence and physical activity adherence. Overall, the predominant best-performing models across all included studies were random forest, decision trees, logistic regression, k-nearest neighbor, LSTM, and ensemble learning.

Model Validation

Owing to the relatively small and imbalanced data sets used in some of the included studies, cross-validation methods were adopted to eliminate bias that may occur during data split. The following cross-validation methods were identified: K(5)-fold cross-validation [29,34]; leave-one-out cross-validation [28,31,33]; and stratified K(10)-fold cross-validation [13,30,35].

Besides cross-validation methods, several performance metrics were used to compare and evaluate the performance of the various machine learning models. It was observed that the choice of performance metrics was dependent on the context of the study and more than 1 metric was used to evaluate the performance of a model (see Table 3). The predominately used metrics in order of frequency included area under the receiver operating characteristic curve (7/11), accuracy (6/11), sensitivity (6/11), specificity (4/11), precision (3/11), F_1 -score (3/11), true

positive rate (3/11), false positive rate (3/11), confusion matrix (3/11), positive predictive value (1/11), negative predictive value (1/11), and precision-recall curve (1/11).

Generally, it was observed that the prediction models had good classification accuracies. This was an indication that the features or predictors used in each of these studies were a good representation of the intervention domain. Nonetheless, due to the plethora of digital platforms, the interaction between technology and behavior may affect the generalizability of the results [33,34]. Thus, Koesmahargyo et al [34] posit replication and integration of data from various digital platforms.

Barriers to Adherence (RQ4)

Studies suggest that the rate of adherence may be affected by the following:

1. Achievement of set health or body goals
2. Issues associated with trust and the tolerability of the technology
3. The complexity of the system, and the mismatch between the system design and the needs and preferences of its users
4. Inappropriate timing for delivering or sending notifications or suggestions to the users; since these timings are usually not well chosen, they may either inconvenience the users when delivered or may not be effective in getting their attention
5. The insufficient open collaborative relationship between health app providers and the users. Prior studies [37] refer to this as a lack of accountability in adherence prediction models

These barriers may be classified into two nonadherent groups: intentional nonadherence (1 and 2), or unintentional nonadherence (3, 4, and 5).

Discussion

This systematic review provides an overview of existing trends in the use of machine learning techniques to predict adherence to different BCSSs. This was achieved by finding answers to a set of review questions using data extracted from the 11 included studies. The rest of this section will summarize and discuss findings based on the review results.

This review identified 4 categories of adherence problems: adherence to digital cognitive or behavioral interventions, medication adherence, physical activity adherence, and diet adherence. These problems collectively represent what Middleton et al [4] describe as an “adherence challenge.” However, when considering the taxonomy of key health behaviors [41], it was observed that the behaviors identified in this systematic review were not exhaustive. Consequently, the prediction of adherence to other health behaviors is an open research area for further investigation.

On the use of persuasive system design features, it was observed that primary task support features were the most used, while social support features were the least used. This finding is consistent with that of related systematic reviews [37]. Though the included studies claimed that either the implementation of the BCSS or the selected features (predictors) for the machine

learning algorithm was theory-based, this systematic review could not identify the behavior change theories adopted by the studies. Hence, it was not clear if the operating mechanisms of behavioral theories were considered in most of the included studies. Nonetheless, prior studies have provided evidence of the effectiveness of theory-based interventions and persuasive system features in promoting adherence behavior in BCSSs. Future studies should therefore be intentional about the use of these mechanisms as measures of improving adherence behavior.

The relevant predictors identified align with findings from existing literature. Similar to existing reviews [4,42], exercise history, intensity, and frequency emerged as relevant predictors of physical activity adherence. Exercise, fatigue, cognitive load, and confidence were the most relevant predictors of diet adherence, affirming previous findings (eg, [42]). While communication with or feedback/advice from the physician or health provider, fear, and patients’ cognitive capacity were the most relevant predictors of medication adherence as also found in past reviews (eg, [3,42]). Furthermore, some of the identified predictors of physical activity adherence and medication adherence affirm 2 viewpoints from social learning theory [43]: that individuals develop beliefs that they can perform the necessary tasks to obtain the desired outcome based on prior accomplishment of similar behaviors and verbal persuasion by credible sources. This systematic review identified the completion of homework assignments as a predictor of cognitive or behavioral intervention adherence, while Heesch et al [44] identified the same predictor for physical activity adherence. Furthermore, this review suggests that not all initially selected predictors or features of adherence are subsequently considered most relevant by the feature selection algorithms (see [Multimedia Appendix 5](#) [14,15,29-37]). Using multiple feature selection methods yields the same feature set. Future studies should consider incorporating the feature selection or engineering techniques identified in this review to enable a comparison of their results with the existing literature.

Most of the included studies used traditional machine learning techniques, with limited use of advanced learning techniques such as ensemble learning, reinforcement learning, and deep learning. Among the 13 supervised machine learning algorithms, only 2 were deep learning techniques (multilayer perceptron and recurrent neural network–LSTM), 1 ensemble, and zero reinforcement learning. Perhaps the sparing use of deep learning techniques may be attributed to the small sample sizes of these studies, considering that deep learning is more efficient in the analysis of huge amounts of data. For instance, LSTM may have been a more appropriate algorithm for the study by Evangelista et al [28], because the data collected captured changes in conditions that evolved slowly over time. However, it was not used probably due to a sample of only 14 participants. Interestingly, in a study with a large data set (342,174 injection historic drop data) [29], LSTM outperformed traditional machine learning models like random forest. Future studies should therefore consider using advanced learning methods instead of traditional learning techniques. Deep learning techniques can automate feature engineering or selection and extract complex and nonlinear patterns from data. Reinforcement learning is well suited for systems with inherent time delays where

decisions are evaluated by a long-term future reward and not an immediate knowledge of the effectiveness of a system [47]. In addition, since they learn by observing the results of their actions, they are applicable in study settings with scarce or varying data as found in BCSSs [22,47].

This systematic review observed that quite a small amount of data were used in most of the included studies. With each study participant treated as a single data point, data were extracted from as little as 12 study participants to as many as 7697 study participants. Although training a machine learning model requires a reasonable amount of data to train the model, the required sample size for training and producing a model with good generalizability is not well established [33,48]. Regardless, the included studies adopted suitable machine learning techniques, dimensionality reduction techniques, and evaluation methods that are designed to improve model performance irrespective of the sample size. For instance, Zhou et al [14] performed data augmentation on the training data.

Multiple metrics can be used to evaluate model performance (see Table 3). However, the choice of which metrics best measure the model performance depends on the nature of the problem, the researcher's understanding of the domain or problem, and the expected outcome of the study. For instance, Gu et al [29] prioritized predicting nonadherers (those who will not perform the recommended task on time), hence specificity (true negative rate) was a preferred metric. Considering that a wrong prediction may lead the health app provider to develop unnecessary persuasive strategies for the user, Pedersen et al [13] and Bastidas et al [27] aimed to reduce false negatives (ie,

participants at high risk of dropout are not identified as such), hence a high precision was a preferred metric. However, if the study's goal is to validate the hypothesis that machine learning methods can be used in predicting adherence [14,31] rather than to compare machine learning methods, then choosing the most appropriate metric becomes irrelevant.

A major research challenge reported in 9 of 11 of the included studies was the scarce and small-sized data sets and their effect on the generalizability and reliability of the research results. A specific study limitation pertained to collecting data in a controlled environment [31]. This method of data collection does not represent the entire range of user behavior in a free-living environment.

Conclusions

Findings from this systematic review indicate that though the use of machine learning techniques in the prediction of adherence to BCSSs is scarce and is only beginning to gain research interest, it has the potential to accurately predict adherence behavior in real time using objectively collected data. This systematic review is unique as it has not yet been reported in the literature, and it provides an overview of machine learning approaches in determining predictors of specific adherence problems in BCSSs. A grasp of these trends across different BCSSs will guide researchers in choosing appropriate features and machine learning techniques that favor the prediction of specific adherence problems. In summarizing findings from 11 journal papers, this systematic review highlights research gaps and areas for future research. It also acknowledges limitations that may exist due to the selection strategy for eligible studies.

Conflicts of Interest

None declared.

Multimedia Appendix 1

Search syntax and results.

[\[DOCX File , 18 KB - ai_v2i1e46779_app1.docx \]](#)

Multimedia Appendix 2

Multimedia quality assessment.

[\[DOCX File , 22 KB - ai_v2i1e46779_app2.docx \]](#)

Multimedia Appendix 3

List of reviewed primary studies.

[\[DOCX File , 19 KB - ai_v2i1e46779_app3.docx \]](#)

Multimedia Appendix 4

Characteristics of the studies and the BCSSs.

[\[DOCX File , 18 KB - ai_v2i1e46779_app4.docx \]](#)

Multimedia Appendix 5

Input data and feature selection approaches.

[\[DOCX File , 31 KB - ai_v2i1e46779_app5.docx \]](#)

References

1. Oinas-Kukkonen H. A foundation for the study of behavior change support systems. *Pers Ubiquit Comput* 2012;17(6):1223-1235. [doi: [10.1007/s00779-012-0591-5](https://doi.org/10.1007/s00779-012-0591-5)]
2. Burgess E, Hassmén P, Pumpa KL. Determinants of adherence to lifestyle intervention in adults with obesity: a systematic review. *Clin Obes* 2017;7(3):123-135. [doi: [10.1111/cob.12183](https://doi.org/10.1111/cob.12183)] [Medline: [28296261](https://pubmed.ncbi.nlm.nih.gov/28296261/)]
3. Aziz F, Malek S, Mhd Ali A, Wong MS, Mosleh M, Milow P. Determining hypertensive patients' beliefs towards medication and associations with medication adherence using machine learning methods. *PeerJ* 2020;8:e8286 [FREE Full text] [doi: [10.7717/peerj.8286](https://doi.org/10.7717/peerj.8286)] [Medline: [32206445](https://pubmed.ncbi.nlm.nih.gov/32206445/)]
4. Middleton KR, Anton SD, Perri MG. Long-term adherence to health behavior change. *Am J Lifestyle Med* 2013;7(6):395-404 [FREE Full text] [doi: [10.1177/1559827613488867](https://doi.org/10.1177/1559827613488867)] [Medline: [27547170](https://pubmed.ncbi.nlm.nih.gov/27547170/)]
5. Aswad O, Lessard L. Service design methods' ability to personalize telehealth: a systematic literature review. 2021 Presented at: The 27th Annual Americas Conference on Information Systems (AMCIS 2021); August 9-13, 2021; Montreal, QC.
6. Abdul-Razak S, Daher AM, Ramli AS, Ariffin F, Mazapuspavina MY, Ambigga KS, et al. Prevalence, awareness, treatment, control and socio demographic determinants of hypertension in Malaysian adults. *BMC Public Health* 2016;16:351 [FREE Full text] [doi: [10.1186/s12889-016-3008-y](https://doi.org/10.1186/s12889-016-3008-y)] [Medline: [27097542](https://pubmed.ncbi.nlm.nih.gov/27097542/)]
7. Patoz M, Hidalgo-Mazzei D, Pereira B, Blanc O, de Chazeron I, Murru A, et al. Patients' adherence to smartphone apps in the management of bipolar disorder: a systematic review. *Int J Bipolar Disord* 2021;9(1):19 [FREE Full text] [doi: [10.1186/s40345-021-00224-6](https://doi.org/10.1186/s40345-021-00224-6)] [Medline: [34081234](https://pubmed.ncbi.nlm.nih.gov/34081234/)]
8. Lie SS, Karlsen B, Oord ER, Graue M, Oftedal B. Dropout from an eHealth intervention for adults with type 2 diabetes: a qualitative study. *J Med Internet Res* 2017;19(5):e187 [FREE Full text] [doi: [10.2196/jmir.7479](https://doi.org/10.2196/jmir.7479)] [Medline: [28559223](https://pubmed.ncbi.nlm.nih.gov/28559223/)]
9. World Health Organization. In: Sabaté E, editor. *Adherence to Long-term Therapies: Evidence for Action*. Geneva, Switzerland: World Health Organization; 2003.
10. Vrijens B, De Geest S, Hughes DA, Przemyslaw K, Demonceau J, Ruppar T, et al. A new taxonomy for describing and defining adherence to medications. *Br J Clin Pharmacol* 2012;73(5):691-705 [FREE Full text] [doi: [10.1111/j.1365-2125.2012.04167.x](https://doi.org/10.1111/j.1365-2125.2012.04167.x)] [Medline: [22486599](https://pubmed.ncbi.nlm.nih.gov/22486599/)]
11. Osterberg L, Blaschke T. Adherence to medication. *N Engl J Med* 2005;353(5):487-497. [doi: [10.1056/NEJMra050100](https://doi.org/10.1056/NEJMra050100)] [Medline: [16079372](https://pubmed.ncbi.nlm.nih.gov/16079372/)]
12. Flynn M, Hall EE. Prediction of adherence to a 9-week corporate wellness walking program. *Health (Irvine Calif)* 2018;10(12):1734-1748 [FREE Full text] [doi: [10.4236/health.2018.1012131](https://doi.org/10.4236/health.2018.1012131)]
13. Pedersen DH, Mansourvar M, Sortsø C, Schmidt T. Predicting dropouts from an electronic health platform for lifestyle interventions: analysis of methods and predictors. *J Med Internet Res* 2019;21(9):e13617 [FREE Full text] [doi: [10.2196/13617](https://doi.org/10.2196/13617)] [Medline: [31486409](https://pubmed.ncbi.nlm.nih.gov/31486409/)]
14. Zhou M, Fukuoka Y, Goldberg K, Vittinghoff E, Aswani A. Applying machine learning to predict future adherence to physical activity programs. *BMC Med Inform Decis Mak* 2019;19(1):169 [FREE Full text] [doi: [10.1186/s12911-019-0890-0](https://doi.org/10.1186/s12911-019-0890-0)] [Medline: [31438926](https://pubmed.ncbi.nlm.nih.gov/31438926/)]
15. Holmes EAF, Hughes DA, Morrison VL. Predicting adherence to medications using health psychology theories: a systematic review of 20 years of empirical research. *Value Health* 2014;17(8):863-876 [FREE Full text] [doi: [10.1016/j.jval.2014.08.2671](https://doi.org/10.1016/j.jval.2014.08.2671)] [Medline: [25498782](https://pubmed.ncbi.nlm.nih.gov/25498782/)]
16. Essery R, Geraghty AWA, Kirby S, Yardley L. Predictors of adherence to home-based physical therapies: a systematic review. *Disabil Rehabil* 2017;39(6):519-534. [doi: [10.3109/09638288.2016.1153160](https://doi.org/10.3109/09638288.2016.1153160)] [Medline: [27097761](https://pubmed.ncbi.nlm.nih.gov/27097761/)]
17. Areerak K, Waongengarm P, Janwantanakul P. Factors associated with exercise adherence to prevent or treat neck and low back pain: a systematic review. *Musculoskelet Sci Pract* 2021;52:102333. [doi: [10.1016/j.msksp.2021.102333](https://doi.org/10.1016/j.msksp.2021.102333)] [Medline: [33529988](https://pubmed.ncbi.nlm.nih.gov/33529988/)]
18. Di Lorito C, Bosco A, Booth V, Goldberg S, Harwood RH, Van der Wardt V. Adherence to exercise interventions in older people with mild cognitive impairment and dementia: a systematic review and meta-analysis. *Prev Med Rep* 2020;19:101139 [FREE Full text] [doi: [10.1016/j.pmedr.2020.101139](https://doi.org/10.1016/j.pmedr.2020.101139)] [Medline: [32793408](https://pubmed.ncbi.nlm.nih.gov/32793408/)]
19. Ormel HL, van der Schoot GGF, Sluiter WJ, Jalving M, Gietema JA, Walenkamp AME. Predictors of adherence to exercise interventions during and after cancer treatment: a systematic review. *Psychooncology* 2018;27(3):713-724 [FREE Full text] [doi: [10.1002/pon.4612](https://doi.org/10.1002/pon.4612)] [Medline: [29247584](https://pubmed.ncbi.nlm.nih.gov/29247584/)]
20. Spanakis G, Weiss G, Boh B, Lemmens L, Roefs A. Machine learning techniques in eating behavior e-coaching. *Pers Ubiquit Comput* 2017;21(4):645-659 [FREE Full text] [doi: [10.1007/s00779-017-1022-4](https://doi.org/10.1007/s00779-017-1022-4)]
21. Triantafyllidis AK, Tsanas A. Applications of machine learning in real-life digital health interventions: review of the literature. *J Med Internet Res* 2019;21(4):e12286 [FREE Full text] [doi: [10.2196/12286](https://doi.org/10.2196/12286)] [Medline: [30950797](https://pubmed.ncbi.nlm.nih.gov/30950797/)]
22. Yom-Tov E, Feraru G, Kozdoba M, Mannor S, Tennenholtz M, Hochberg I. Encouraging physical activity in patients with diabetes: intervention using a reinforcement learning system. *J Med Internet Res* 2017;19(10):e338 [FREE Full text] [doi: [10.2196/jmir.7994](https://doi.org/10.2196/jmir.7994)] [Medline: [29017988](https://pubmed.ncbi.nlm.nih.gov/29017988/)]
23. Bohlmann A, Mostafa J, Kumar M. Machine learning and medication adherence: scoping review. *JMIRx Med* 2021;2(4):e26993 [FREE Full text] [doi: [10.2196/26993](https://doi.org/10.2196/26993)] [Medline: [37725549](https://pubmed.ncbi.nlm.nih.gov/37725549/)]
24. Falagas ME, Pitsouni EI, Malietzis GA, Pappas G. Comparison of PubMed, Scopus, Web of Science, and Google Scholar: strengths and weaknesses. *FASEB J* 2008;22(2):338-342 [FREE Full text] [doi: [10.1096/fj.07-9492LSF](https://doi.org/10.1096/fj.07-9492LSF)] [Medline: [17884971](https://pubmed.ncbi.nlm.nih.gov/17884971/)]

25. Aromataris E, Fernandez R, Godfrey CM, Holly C, Khalil H, Tungpunkom P. Summarizing systematic reviews: methodological development, conduct and reporting of an umbrella review approach. *Int J Evid Based Healthc* 2015;13(3):132-140 [[FREE Full text](#)] [doi: [10.1097/XEB.000000000000055](https://doi.org/10.1097/XEB.000000000000055)] [Medline: [26360830](#)]
26. Hong QN, Fàbregues S, Bartlett G, Boardman F, Cargo M, Dagenais P, et al. The Mixed Methods Appraisal Tool (MMAT) version 2018 for information professionals and researchers. *Educ Inf* 2018;34(4):285-291. [doi: [10.3233/EFI-180221](https://doi.org/10.3233/EFI-180221)]
27. Jossa-Bastidas O, Zahia S, Fuente-Vidal A, Férez NS, Noguera OR, Montane J, et al. Predicting physical exercise adherence in fitness apps using a deep learning approach. *Int J Environ Res Public Health* 2021;18(20):10769 [[FREE Full text](#)] [doi: [10.3390/ijerph182010769](https://doi.org/10.3390/ijerph182010769)] [Medline: [34682515](#)]
28. Evangelista LS, Ghasemzadeh H, Lee JA, Fallahzadeh R, Sarrafzadeh M, Moser DK. Predicting adherence to use of remote health monitoring systems in a cohort of patients with chronic heart failure. *Technol Health Care* 2017;25(3):425-433 [[FREE Full text](#)] [doi: [10.3233/THC-161279](https://doi.org/10.3233/THC-161279)] [Medline: [27886024](#)]
29. Gu Y, Zalkikar A, Liu M, Kelly L, Hall A, Daly K, et al. Predicting medication adherence using ensemble learning and deep learning models with large scale healthcare data. *Sci Rep* 2021;11(1):18961 [[FREE Full text](#)] [doi: [10.1038/s41598-021-98387-w](https://doi.org/10.1038/s41598-021-98387-w)] [Medline: [34556746](#)]
30. Ramos LA, Blankers M, van Wingen G, de Bruijn T, Pauws SC, Goudriaan AE. Predicting success of a digital self-help intervention for alcohol and substance use with machine learning. *Front Psychol* 2021;12:734633 [[FREE Full text](#)] [doi: [10.3389/fpsyg.2021.734633](https://doi.org/10.3389/fpsyg.2021.734633)] [Medline: [34552539](#)]
31. Tucker CS, Behoora I, Nembhard HB, Lewis M, Sterling NW, Huang X. Machine learning classification of medication adherence in patients with movement disorders using non-wearable sensors. *Comput Biol Med* 2015;66:120-134 [[FREE Full text](#)] [doi: [10.1016/j.combiomed.2015.08.012](https://doi.org/10.1016/j.combiomed.2015.08.012)] [Medline: [26406881](#)]
32. Wallert J, Gustafson E, Held C, Madison G, Norlund F, von Essen L, et al. Predicting adherence to internet-delivered psychotherapy for symptoms of depression and anxiety after myocardial infarction: machine learning insights from the U-CARE heart randomized controlled trial. *J Med Internet Res* 2018;20(10):e10754 [[FREE Full text](#)] [doi: [10.2196/10754](https://doi.org/10.2196/10754)] [Medline: [30305255](#)]
33. Goldstein SP, Zhang F, Thomas JG, Butryn ML, Herbert JD, Forman EM. Application of machine learning to predict dietary lapses during weight loss. *J Diabetes Sci Technol* 2018;12(5):1045-1052 [[FREE Full text](#)] [doi: [10.1177/1932296818775757](https://doi.org/10.1177/1932296818775757)] [Medline: [29792067](#)]
34. Koesmahargyo V, Abbas A, Zhang L, Guan L, Feng S, Yadav V, et al. Accuracy of machine learning-based prediction of medication adherence in clinical research. *Psychiatry Res* 2020;294:113558 [[FREE Full text](#)] [doi: [10.1016/j.psychres.2020.113558](https://doi.org/10.1016/j.psychres.2020.113558)] [Medline: [33242836](#)]
35. Bremer V, Chow PI, Funk B, Thorndike FP, Ritterband LM. Developing a process for the analysis of user journeys and the prediction of dropout in digital health interventions: machine learning approach. *J Med Internet Res* 2020;22(10):e17738 [[FREE Full text](#)] [doi: [10.2196/17738](https://doi.org/10.2196/17738)] [Medline: [33112241](#)]
36. Nasseh K, Frazee SG, Visaria J, Vlahiotis A, Tian Y. Cost of medication nonadherence associated with diabetes, hypertension, and dyslipidemia. *Am J Pharm Benefits* 2012;4(2):e41-e47.
37. Ekpezu A, Wiafe I, Oinas-kukkonen H. Technological factors that influence user compliance with behavior change support systems: a systematic review. *Proc Annu Hawaii Int Conf Syst Sci Hawaii* . [doi: [10.2196/preprints.46779](https://doi.org/10.2196/preprints.46779)]
38. Nutrokpor C, Ekpezu A, Wiafe A, Wiafe I. Exploring the impact of persuasive system features on user sentiments in health and fitness apps. 2021 Presented at: Proceedings of the Ninth International Workshop on Behavior Change Support Systems, BCSS 2021; April 12-14, 2021; Bournemouth, UK p. 25-39.
39. Irshad MT, Nisar MA, Huang X, Hartz J, Flak O, Li F, et al. SenseHunger: machine learning approach to hunger detection using wearable sensors. *Sensors (Basel)* 2022;22(20):7711 [[FREE Full text](#)] [doi: [10.3390/s22207711](https://doi.org/10.3390/s22207711)] [Medline: [36298061](#)]
40. Tang J, Alelyani S, Liu H. Feature selection for classification: a review. In: Aggarwal CC, editor. *Data Classification: Algorithms and Applications*. New York, NY: CRC Press; 2014:37-64.
41. Ratz T, Lippke S. 8.06—health behavior change. *Clin Psychol Sci (Second Ed)* 2022;8:95-117. [doi: [10.1016/b978-0-12-818697-8.00068-6](https://doi.org/10.1016/b978-0-12-818697-8.00068-6)]
42. Culos-Reed SN, Rejeski WJ, McAuley E, Ockene JK, Roter DL. Predictors of adherence to behavior change interventions in the elderly. *Control Clin Trials* 2000;21(5 Suppl):200S-205S. [doi: [10.1016/s0197-2456\(00\)00079-9](https://doi.org/10.1016/s0197-2456(00)00079-9)] [Medline: [11018576](#)]
43. Bandura A. *Social Foundations of Thought and Action: A Social Cognitive Theory*. Englewood Cliffs, NJ: Prentice-Hall; 1986.
44. Heesch KC, Mâsse LC, Dunn AL, Frankowski RF, Mullen PD. Does adherence to a lifestyle physical activity intervention predict changes in physical activity? *J Behav Med* 2003;26(4):333-348. [doi: [10.1023/a:1024205011001](https://doi.org/10.1023/a:1024205011001)] [Medline: [12921007](#)]
45. Oinas-Kukkonen H, Harjumaa M. Persuasive systems design: key issues, process model, and system features. *Commun Assoc Inf Syst* 2009;24(1):485-500. [doi: [10.17705/ICAIS.02428](https://doi.org/10.17705/ICAIS.02428)]
46. Merz M, Augsburg U, Ackermann L, Ackermann L. Design principles of persuasive systems—review and discussion of the persuasive systems design model. 2021 Presented at: AMCIS 2021 Proceedings; August 9-13, 2021; Montreal, Canada URL: https://aisel.aisnet.org/amcis2021/sig_hci/sig_hci/3

47. Yu C, Liu J, Nemati S, Yin G. Reinforcement learning in healthcare: a survey. *ACM Comput Surv* 2021;55(1):1-36 [[FREE Full text](#)] [doi: [10.1145/3477600](https://doi.org/10.1145/3477600)]
48. Balki I, Amirabadi A, Levman J, Martel AL, Emersic Z, Meden B, et al. Sample-size determination methodologies for machine learning in medical imaging research: a systematic review. *Can Assoc Radiol J* 2019 Nov;70(4):344-353 [[FREE Full text](#)] [doi: [10.1016/j.carj.2019.06.002](https://doi.org/10.1016/j.carj.2019.06.002)] [Medline: [31522841](https://pubmed.ncbi.nlm.nih.gov/31522841/)]

Abbreviations

BCSS: behavior change support system

LSTM: long short-term memory

MI-ANXDEP: myocardial infarction-anxiety, depression, or both

PSD: persuasive systems design

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

RQ: review question

Edited by K El Emam, B Malin; submitted 02.03.23; peer-reviewed by R Marshall, C Manlhiot; comments to author 31.07.23; revised version received 20.09.23; accepted 28.10.23; published 22.11.23.

Please cite as:

Ekpezu AO, Wiafe I, Oinas-Kukkonen H

Predicting Adherence to Behavior Change Support Systems Using Machine Learning: Systematic Review

JMIR AI 2023;2:e46779

URL: <https://ai.jmir.org/2023/1/e46779>

doi: [10.2196/46779](https://doi.org/10.2196/46779)

PMID:

©Akon Obu Ekpezu, Isaac Wiafe, Harri Oinas-Kukkonen. Originally published in JMIR AI (<https://ai.jmir.org>), 22.11.2023. This is an open-access article distributed under the terms of the Creative Commons Attribution License (<https://creativecommons.org/licenses/by/4.0/>), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR AI, is properly cited. The complete bibliographic information, a link to the original publication on <https://www.ai.jmir.org/>, as well as this copyright and license information must be included.

Review

Machine Learning–Based Asthma Attack Prediction Models From Routinely Collected Electronic Health Records: Systematic Scoping Review

Arif Budiarto^{1,2}, MSc; Kevin C H Tsang¹, PhD; Andrew M Wilson^{3,4}, MD FRCP; Aziz Sheikh¹, MD; Syed Ahmar Shah¹, DPhil

¹Asthma UK Center for Applied Research, Usher Institute, University of Edinburgh, Edinburgh, United Kingdom

²Bioinformatics and Data Science Research Center, Bina Nusantara University, Jakarta, Indonesia

³Norwich Medical School, University of East Anglia, Norwich, United Kingdom

⁴Norfolk and Norwich University Hospital NHS Foundation Trust, Norwich, United Kingdom

Corresponding Author:

Arif Budiarto, MSc

Asthma UK Center for Applied Research

Usher Institute

University of Edinburgh

NINE, 9 Little France Road

Edinburgh BioQuarter

Edinburgh, EH16 4UX

United Kingdom

Phone: 44 7447900766

Email: arif.budiarto@ed.ac.uk

Abstract

Background: An early warning tool to predict attacks could enhance asthma management and reduce the likelihood of serious consequences. Electronic health records (EHRs) providing access to historical data about patients with asthma coupled with machine learning (ML) provide an opportunity to develop such a tool. Several studies have developed ML-based tools to predict asthma attacks.

Objective: This study aims to critically evaluate ML-based models derived using EHRs for the prediction of asthma attacks.

Methods: We systematically searched PubMed and Scopus (the search period was between January 1, 2012, and January 31, 2023) for papers meeting the following inclusion criteria: (1) used EHR data as the main data source, (2) used asthma attack as the outcome, and (3) compared ML-based prediction models' performance. We excluded non-English papers and nonresearch papers, such as commentary and systematic review papers. In addition, we also excluded papers that did not provide any details about the respective ML approach and its result, including protocol papers. The selected studies were then summarized across multiple dimensions including data preprocessing methods, ML algorithms, model validation, model explainability, and model implementation.

Results: Overall, 17 papers were included at the end of the selection process. There was considerable heterogeneity in how asthma attacks were defined. Of the 17 studies, 8 (47%) studies used routinely collected data both from primary care and secondary care practices together. Extreme imbalanced data was a notable issue in most studies (13/17, 76%), but only 38% (5/13) of them explicitly dealt with it in their data preprocessing pipeline. The gradient boosting–based method was the best ML method in 59% (10/17) of the studies. Of the 17 studies, 14 (82%) studies used a model explanation method to identify the most important predictors. None of the studies followed the standard reporting guidelines, and none were prospectively validated.

Conclusions: Our review indicates that this research field is still underdeveloped, given the limited body of evidence, heterogeneity of methods, lack of external validation, and suboptimally reported models. We highlighted several technical challenges (class imbalance, external validation, model explanation, and adherence to reporting guidelines to aid reproducibility) that need to be addressed to make progress toward clinical adoption.

(JMIR AI 2023;2:e46717) doi:[10.2196/46717](https://doi.org/10.2196/46717)

KEYWORDS

asthma attack; exacerbation; prognosis; machine learning; electronic health record; review; EHR; asthma

Introduction**Background**

Asthma is a chronic lung illness characterized by reversible airway blockage caused by inflammation and narrowing of the small airways in the lungs that can lead to cough, wheezing, chest tightness, and breathing difficulties [1]. It is a common noncommunicable disease that affects children and adults alike. In 2019, asthma affected an estimated 262 million individuals, resulting in 461,000 fatalities [1,2]. Asthma attacks occur particularly in those with poorly controlled diseases [3]. An asthma attack is a sudden or gradual deterioration of asthma symptoms that can have a major influence on a patient's quality of life [4]. Such attacks can be life-threatening and necessitate rapid medical attention, such as an accident and emergency department visit or hospitalization, and can even lead to mortality [5]. Asthma attacks are prevalent, with >90,000 annual hospital admissions in the United Kingdom alone [6]. Early warning tools to predict asthma attacks offer the opportunity to provide timely treatments and, thereby, minimize the risk of serious outcomes [4].

Machine learning (ML) offers the potential to develop an early warning tool that takes different risk factors as input and then outputs the probability of an adverse outcome. So far, logistic regression (LR) has been the most common approach in building an asthma attack risk prediction tool [7-9]. However, the predictive performance of this method may be inferior to more advanced ML methods, especially for relatively high-dimensional data with complex and nonlinear relationships between the variables [10,11]. The use of ML has been investigated in a wide range of medical domains by using various data such as electronic health records (EHRs), medical images, genomics data, and wearables data [12-14]. However, to the best of our knowledge, there is still no widely used ML-based asthma attack risk prediction tool in clinical practice.

Objective

Previous recent systematic reviews have discussed the choice of models used for asthma prognosis [15,16]. An ML pipeline, however, has several components besides modeling choice (eg, feature engineering [17]), which can profoundly influence the performance of the algorithms. Owing to the lack of consensus about what constitutes best practices for the application of ML for predicting asthma attacks, there is considerable heterogeneity in previous studies [15,16], thereby making direct comparisons challenging. In this scoping review, we aimed to critically examine existing studies that used ML algorithms for the prediction of asthma attacks with routinely collected EHR data. Besides data type and choice of models, we have reviewed additional ML pipeline challenges. These include customizing *off-the-shelf* algorithms to account for domain-specific subtleties and the need for the model to be explainable, extensively validated (externally and prospectively), and transparently reported.

Methods**Overview**

The scoping review was conducted based on the 5-stage framework by Arksey and O'Malley [18]. This framework includes identifying the research question; searching and collecting relevant studies; study filtering; data charting; and finally, collating, summarizing, and reporting the results. The research questions in this scoping review were the following:

1. What methods are commonly used in developing an asthma attack prediction model?
2. How did the authors process the features and outcome variables?
3. Are there any of these prediction models that have been implemented in a real-world clinical setting?

We then translated these 3 questions into the patient or population, intervention, comparison, and outcomes model [19,20], as shown in Table 1.

Table 1. The patient or population, intervention, comparison, and outcomes structure.

Item	Expansion	Keywords
P	Patient, population	People with asthma
I	Intervention, prognostic factor, or exposure	Machine learning
C	Comparison of intervention	N/A ^a
O	Outcome	Asthma attack

^aN/A: not applicable.

Search Strategy

We used the patient or population, intervention, comparison, and outcomes model in Table 1 as our framework for defining relevant keywords. This approach led us to include clinical terms associated with asthma attacks, encompassing concepts such as asthma exacerbation, asthma control, asthma

management, and hospitalization. In addition, we integrated technical terminology related to ML, incorporating terms such as artificial intelligence, supervised methods, and deep learning (DL). All the keywords that we used in the search strategy can be found in Multimedia Appendix 1 [4,11,21-35]. Overall, 2 databases, PubMed and Scopus, were chosen as the sources of papers. The search period was between January 1, 2012, and

January 31, 2023, and the search was limited to the title, abstract, and keywords of each paper but without any language restriction. The complete query syntax for both databases is listed in [Textbox 1](#).

Textbox 1. Query syntax.

Scopus

- ((TITLE-ABS-KEY("asthma") AND (TITLE-ABS-KEY("management") OR TITLE-ABS-KEY("control") OR TITLE-ABS-KEY("attack") OR TITLE-ABS-KEY("exacerbation") OR TITLE-ABS-KEY("risk stratification") OR TITLE-ABS-KEY("risk prediction") OR TITLE-ABS-KEY("risk classification") OR TITLE-ABS-KEY("hospitalization") OR TITLE-ABS-KEY("hospitalisation") OR TITLE-ABS-KEY("prognosis")))) AND (TITLE-ABS-KEY("machine learning") OR TITLE-ABS-KEY("artificial intelligence") OR TITLE-ABS-KEY("supervised method") OR TITLE-ABS-KEY("unsupervised method") OR TITLE-ABS-KEY("deep learning") OR TITLE-ABS-KEY("supervised learning") OR TITLE-ABS-KEY("unsupervised learning")))) AND PUBYEAR > 2011

PubMed

- ((asthma[Text Word]) AND ((Management[Text Word]) OR (Control[Text Word]) OR (Attack[Text Word]) OR (Exacerbation[Text Word]) OR (Risk Stratification[Text Word]) OR (Risk Prediction[Text Word]) OR (Risk Classification[Text Word]) OR (hospitalization[Text Word]) OR (hospitalisation[Text Word]) OR (prognosis[Text Word])) AND ((machine learning[Text Word]) OR (Artificial Intelligence[Text Word]) OR (supervised method[Text Word]) OR (unsupervised method[Text Word]) OR (deep learning[Text Word]) OR (supervised learning[Text Word]) OR (unsupervised learning[Text Word])))) AND ("2012/01/01"[Date - Publication] : "2023/01/31"[Date - Publication])

Eligibility Criteria and Study Selection

Overall, 2 authors (AB and KCHT) performed the 2-step study selection process. During the first selection step, we focused on the abstract. In the second step, we conducted a thorough reading of the full text of the manuscript. We only included papers that met our inclusion criteria: (1) used asthma attack as the outcome, (2) included an ML-based prediction model, and (3) used EHR data as the main data source. We defined the concept of EHR-derived data as structured, text-based, individual-level, and routinely collected data gathered within the health care system. In cases of unclear information extracted from the abstract, the reviewers decided to retain the studies for the next iteration (full-text review). We excluded nonresearch papers, such as commentary and systematic review papers because of the insufficient technical information. We also filtered out papers that did not provide sufficient details about the ML approach and the result, including protocol papers.

Data Extraction

From each of the eligible papers, we extracted data from the full text and web-based supplements. We then summarized these data under different categories such as data set (whether publicly available or not), population characteristics (source, size, age range, and region), year of data, outcome definition and how it was represented in the model, number of features, feature selection method, imbalance handling strategy, ML prediction methods, performance evaluation metric, evaluation result, external validation, explainability method, and real-world clinical setting implementation. The data extraction and

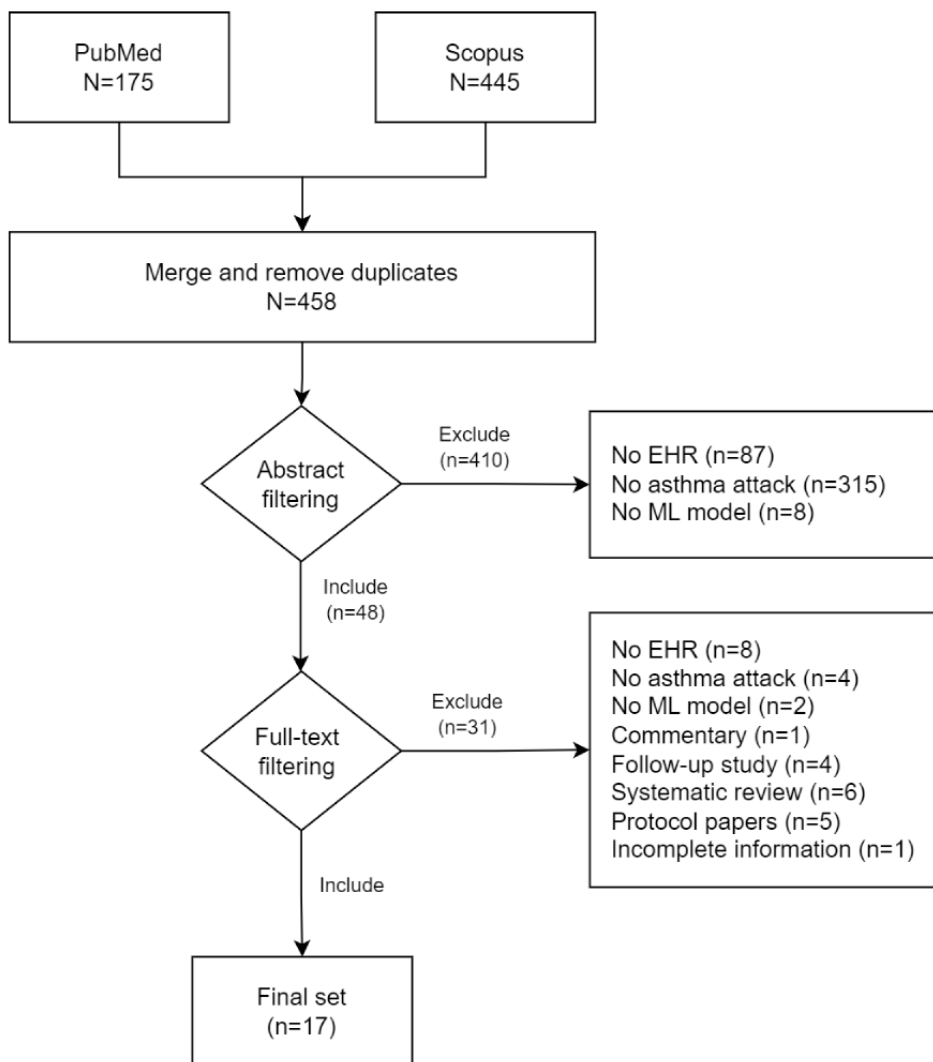
summarization for each paper were conducted independently by 2 authors (AB and KCHT). In case of any discrepancies, the 2 authors discussed them in detail during face-to-face meetings to reach an agreement. If the 2 reviewers could not resolve the disagreement, we had a further discussion with the whole team. For each study, we have reported both the performance evaluation result of the prediction models and the most important predictors where available.

Results

Overview

In total, 458 nonduplicated, potentially eligible papers were identified. At the end of the selection process, 3.7% (17/458) of the papers were included based on the inclusion criteria (refer to the PRISMA [Preferred Reporting Items for Systematic Reviews and Meta-Analyses] diagram in [Figure 1](#)). The earliest study that was included in the full review was published in 2018. In the abstract filtering stage, most of the studies (353/458, 77.1%) were excluded because the prediction outcome was not an asthma attack. We included 10.5% (48/458) of the studies in the full-text filtering stage. Eventually, 3.1% (14/458) of the studies were excluded because they did not meet our inclusion criteria. Then, 2.6% (12/458) nonresearch papers were also excluded. In addition, we excluded 0.9% (4/458) of the studies, which were a follow-up for 2 main papers that we included in the extraction stage. All the summary points in these follow-up studies were identical to the ones in the main studies. We also excluded 0.2% (1/458) of the studies owing to insufficient information.

Figure 1. PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) diagram. EHR: electronic health record; ML: machine learning.



Asthma Data Sets

Table 2 summarizes the basic information about each included study. Only 6% (1/17) of the studies used routinely collected data from primary care alone [21]. Of the 17 studies, only 8 (47%) used data from secondary care, and the remaining 8 (47%) used routinely collected data from both primary and secondary care. All studies used data sets hosted either at the author’s institution or their collaborators’ institution, except a study [22] that used publicly available data (the Medical Information Mart for Intensive Care III data set [36]) as one of their data sets. Overall, 76% (13/17) of the studies used only EHR data to build the prediction model. Of the 17 studies, 4 (24%) studies

integrated EHR data with additional modalities, including radiology images (chest computed tomography scans) [23] and environmental data [11,24,25], aiming to enhance predictive accuracy. The study populations varied across the studies, with most of them involving adults (8/17, 47%), followed by the general population, both children and adults (5/17, 29%), and children (4/17, 24%). Of the 17 studies, 14 (82%) had study populations from the United States. The other countries studied included Japan, Sweden, and the United Kingdom. All studies incorporated >1000 samples, except a study [23] that trained the prediction model on <200 samples. Among the studies, the biggest data set had data from 397,858 patients [26].

Table 2. Summary of studies' basic information.

Study, year	Health care setting	Publicly available data set	Data source	Region	Data year	Sample size
Inselman et al [27], 2022	Secondary care	No	Single modality	United States	2003-2020	3057
Hurst et al [25], 2022	Both	No	Multimodality	United States	2014-2019	5982
Hogan et al [28], 2022	Secondary care	No	Single modality	United States	2013	18,489
Zein et al [29], 2021	Both	No	Single modality	United States	2010-2018	60,302
Sills et al [30], 2021	Secondary care	No	Single modality	United States	2009-2013	9069
Hozawa et al [31], 2021	Secondary care	No	Single modality	Japan	2016-2017	42,685
Lisspers et al [32], 2021	Both	No	Single modality	Sweden	2000-2013	29,396
Ananth et al [23], 2021	Secondary care	No	Multimodality	United Kingdom	2018-2020	200
Tong et al [33], 2021	Both	No	Single modality	United States	2011-2018	82,888
Mehrish et al [24], 2021	Secondary care	No	Multimodality	United States	2013-2017	10,000
Xiang et al [4], 2020	Both	No	Single modality	United States	1992-2015	31,433
Cobian et al [34], 2020	Both	No	Single modality	United States	2007-2011	28,101
Luo et al [35], 2020	Both	No	Single modality	United States	2005-2018	315,308
Roe et al [22], 2020	Secondary care	Yes	Single modality	United States	2001-2012	38,597
Luo et al [26], 2020	Both	No	Single modality	United States	2012-2018	397,858
Wu et al [21], 2018	Primary care	No	Single modality	United States	1997-2002	4013
Patel et al [11], 2018	Secondary care	No	Multimodality	United States	2012-2015	29,392

Data Preprocessing

There was considerable heterogeneity in the definition of the prediction outcome used in the models, including asthma exacerbation [4,25,27,29,31,32,34], asthma-related hospitalization [11,24,26,30,33,35], asthma readmission [28], asthma prevalence [24], asthma-related mortality [22], and asthma relapse [21].

The time horizon used to define the prediction outcome also varied across studies. Of the 17 studies, 6 (35%) studies defined the model task as a 1-year prediction [4,23,26,31,33,35]. Other variations in the time horizon for the outcome were 180 days [28], 90 days [34], 28 days [29], and 15 days [32]. A study compared the prediction model performances across 3 time horizons: 30, 90, and 180 days [25]. Of the 17 studies, 2 (12%) studies undertook a different approach, where the aim was to predict asthma attack-related hospitalization within 2 hours after an accident and emergency department visit [11,30]. Of the 17 studies, 3 (18%) studies did not report the prediction time horizon [21,22,24].

There was an obvious class imbalance in 76% (13/17) of the studies (Table 3). Class imbalance is a problem where the distribution of samples across the classes is skewed [37]. Ignoring this problem during model development will produce a biased model. Among the selected studies, the smallest minority class ratio accounted for as little as 0.04% [32]. Among these 17 studies, only 5 (29%) [4,21,30,32,33] explicitly mentioned their strategies to appropriately handle imbalanced data. Synthetic minority oversampling technique [38], oversampling [39,40], and undersampling [39,40] were the

methods reported in these studies. The objective of these 3 methods is to balance the proportion of samples in each class by either generating synthetic data from the minority class or omitting a certain number of samples in the majority class. Of the 17 studies, only 2 (12%) studies used a balanced data set [22,23], whereas 2 (12%) other studies did not report the class ratio of their data set [24,34]. Various feature selection methods were explicitly mentioned as part of the data preprocessing step, including backward stepwise variable selection [28], light gradient boosting method feature importance [32], and Pearson correlation [32]. Of the 17 studies, 5 (29%) studies [4,26,30,33,35] implemented the feature selection process as the built-in method in the model development phase, whereas the remaining studies did not mention the feature selection method in their report. The smallest feature set used in the study was 7 variables [24], and the biggest set was >500 variables [32]. The handling of missing values varied across the studies. In most cases (9/17, 53%), missing values were treated either as a distinct category or assigned a specific value [21,23,25-27,29,32,33,35]. However, some studies opted to exclude data containing missing values [4,11,28,30], whereas others did not specify their approach for addressing this issue [22,24,31,34]. Notably, more than half of the studies (11/17, 65%) did not describe their methods for data normalization. This step is particularly critical for certain ML algorithms such as LR and support vector machine to prevent uneven weighting of features in the model. In contrast, 35% (6/17) of the studies [11,22,23,26,33,35] used a standard mean normalization technique to standardize the continuous features, ensuring uniform scaling across the data set.

Table 3. Summary of the data preprocessing step.

Study, year	Outcomes	Prediction time horizon	Class imbalance ratio (%)	Data imbalance handling methods	Feature selection methods	Number of features
Inselman et al [27], 2022	Asthma exacerbation	180 d	• 22.60	Unknown	Unknown	21
Hurst et al [25], 2022	Asthma exacerbation	30, 90, and 180 d	• 37	Unknown	Unknown	41
Hogan et al [28], 2022	Asthma readmission	180 d	• 5.70	Unknown	Backward step-wise variable selection	21
Zein et al [29], 2021	Asthma exacerbation	28 d	• Nonsevere=32.80 • Severe=2.90	Unknown	Unknown	82
Sills et al [30], 2021	Asthma-related hospitalization	Admission after A&E ^a department visit	• 22.50	Oversampling	Automated feature selection	13
Hozawa et al [31], 2021	Asthma exacerbation	365 d	• 13.70	Unknown	Unknown	25
Lisspers et al [32], 2021	Asthma exacerbation	15 d	• 0.04	Undersampling and weighting method	Correlation and LGBM ^b model	>500
Ananth et al [23], 2021	Asthma exacerbation	365 d	• 50	Unknown	Unknown	17
Tong et al [33], 2021	Asthma-related hospitalization or A&E department visit	365 d	• 1.66	WEKA ^c	Automated feature selection	234
Mehrish et al [24], 2021	Asthma prevalence, asthma-related hospitalization, or hospital readmission	Unknown	• Unknown	Unknown	Unknown	7
Xiang et al [4], 2020	Asthma exacerbation	365 d	• 7.20	SMOTE ^d	Automated feature selection	Unknown
Cobian et al [34], 2020	Asthma exacerbation	90 d	• Unknown	Unknown	Unknown	>25
Luo et al [35], 2020	Asthma-related hospitalization	365 d	• 3.59	Unknown	Automated feature selection	235
Roe et al [22], 2020	Asthma-related mortality	Unknown	• 49	Unknown	Unknown	42
Luo et al [26], 2020	Asthma-related hospitalization	365 d	• 2.30	Unknown	Automated feature selection	337
Wu et al [21], 2018	Asthma relapse	Unknown	• 32.89	Random undersampling	Unknown	60
Patel et al [11], 2018	Asthma-related hospitalization	Admission after ED ^e visit	• 17	Unknown	Unknown	100

^aA&E: accident and emergency.

^bLGBM: light gradient boosting method.

^cWEKA: Waikato Environment for Knowledge Analysis.

^dSMOTE: synthetic minority oversampling technique.

^eED: emergency department.

ML Methods and Performance Evaluation

Table 4 describes the ML and performance evaluation methods used in the selected studies. We found a wide range of ML methods in the selected studies. Most (14/17, 82%) used conventional ML methods such as support vector machine [41], random forest [42], naïve Bayes [43], decision tree (DT) [44],

K-nearest neighbor [45], and artificial neural network [46]. LR and its variations (ie, Ridge, Lasso, and Elastic Net) [47] were found to be the most common baseline model among the studies (10/15, 67%) [4,11,22-25,27-30,32,34]. Some studies developed the prediction model with more advanced ML algorithms such as gradient boosting DT (GBDT)-based methods

[11,22,25-27,29,31-33,35] and DL-based methods [4,21,34]. A few studies [26,30,35] also used automated model selection tools, such as Waikato Environment for Knowledge Analysis [48] and autoML [49]. GBDT-based methods including extreme gradient boosting (XGBoost) [50] were the common best-performing models (area under the curve scores ranging from 0.6 to 0.9). The model performances in all studies were evaluated using the area under the receiver operating characteristic curve score, except in a study [21] that used F_1 -score as the only performance metric. Half of them (9/17, 53%) included additional evaluation metrics such as accuracy,

precision, recall, sensitivity, specificity, positive predictive value, negative predictive value, F_1 -score, area under the precision-recall curve, and microaveraged accuracy [21,23,25-27,30,32,33,35]. Owing to different data sets and the heterogeneity in the definitions of the outcome, prediction time horizon, and preprocessing across the studies, we considered a direct comparison across studies based on the reported evaluation metric to be inappropriate. Only 18% (3/17) of the studies included external validation using retrospective studies in their analysis pipeline [21,26,33].

Table 4. Summary of machine learning (ML) methods.

Study, year	ML methods	Best models	Best performance metrics	External validation
Inselman et al [27], 2022	GLMNet ^a , RF ^b , and GBM ^c	GBM	• AUC ^d =0.74	No
Hurst et al [25], 2022	Lasso, RF, and XGBoost ^e	XGBoost	• 30-d AUC=0.761 • 90-d AUC=0.752 • 180-d AUC=0.739	No
Hogan et al [28], 2022	Cox proportional hazard, LR ^f , and ANN ^g	ANN	• AUC=0.636	No
Zein et al [29], 2021	LR, RF, and GBDT ^h	GBDT	• Nonsevere AUC=0.71 • Hospitalization AUC=0.85 • ED ⁱ AUC=0.88	No
Sills et al [30], 2021	AutoML, RF, and LR	AutoML	• AUC=0.914	No
Hozawa et al [31], 2021	XGBoost	XGBoost	• AUC=0.656	No
Lisspers et al [32], 2021	XGBoost, LGBM ^j , RNN ^k , and LR (Lasso, Ridge, and Elastic Net)	XGBoost	• AUC=0.90	No
Ananth et al [23], 2021	LR, DT ^l , and ANN	LR	• AUC=0.802	No
Tong et al [33], 2021	WEKA ^m and XGBoost	XGBoost	• AUC=0.902	Yes
Mehrish et al [24], 2021	GLM ⁿ , correlation models, and LR	LR	• AUC=0.78	No
Xiang et al [4], 2020	LR, MLP ^o , and LSTM ^p with an attention mechanism	LSTM with an attention mechanism	• AUC=0.7003	No
Cobian et al [34], 2020	LR, RF, and LSTM	LR with L1 (Ridge)	• AUC=0.7697	No
Luo et al [35], 2020	WEKA and XGBoost	XGBoost	• AUC=0.859	No
Roe et al [22], 2020	XGBoost, NN ^q , LR, and KNN ^r	XGBoost	• AUC=0.75	No
Luo et al [26], 2020	WEKA and XGBoost	XGBoost	• AUC=0.820	Yes
Wu et al [21], 2018	LSTM	LSTM	• Binary classification F1-score=0.8508 • Multiclass classification F1-score=0.4976	Yes
Patel et al [11], 2018	DT, Lasso, RF, and GBDT	GBDT	• AUC=0.84	No

^aGLMNet: Lasso and Elastic-Net Regularized Generalized Linear Models.

^bRF: Random Forest.

^cGBM: gradient boosting method.

^dAUC: area under the curve.

^eXGBoost: extreme gradient boosting.

^fLR: logistic regression.

^gANN: artificial neural network.

^hGBDT: gradient boosting decision tree.

ⁱED: emergency department.

^jLGBM: light gradient boosting method.

^kRNN: recurrent neural network.

^lDT: decision tree.

^mWEKA: Waikato Environment for Knowledge Analysis.

ⁿGLM: Generalized Linear Model.

^oMLP: multilayers perceptron.

^pLSTM: long short-term memory.

^qNN: neural network.

^rKNN: K-nearest neighbor.

Model Explainability and Implementation

We then compared how model explainability was handled across studies. Model explainability refers to the degree of transparency and the level of detail a model can provide to offer additional information about its output, facilitating a better understanding of how the model operates [51]. We grouped the studies into 2 categories based on their best model's transparency. In the first group, we included 18% (3/17) of the studies in which the best-performing model can be considered as a transparent model [51], including LR [23,24,34]. However, only 67% (2/3) of them provided a report on this model explanation in the form of LR coefficient values for each variable [23,34]. We grouped the remaining studies into an opaque model category where a post hoc analysis was needed to explain the model prediction mechanism [51]. In this group, all studies [4,11,22,26,28-33,35] used a model-specific method for explaining the prediction mechanism, except for 14% (2/14) of the studies [27,29] that used a model-agnostic method called the shapely additive explanation (SHAP) method [29]. Overall, 14% (2/14) of the studies in this group did not include any model explanation approach [21,25]. Although model-specific explanation methods, such as those used in DT-based models, gauge the impact of each feature on a model's decision through specific metrics developed during training, the SHAP method takes a more comprehensive approach. SHAP conducts a deductive assessment by exploring all the potential combinations of

features to determine how each one influences the final prediction.

None of the studies followed any specific reporting guidelines. Furthermore, despite promising performances in some studies, none were implemented in a real-world clinical setting for prospective evaluation. In each of the studies reviewed, various limitations were identified, encompassing both clinical and nonclinical factors. One of the common limitations in these studies was the issue of generalizing their findings to different health care settings and patient groups [22,25,26,29,33,35]. This difficulty often arose because they lacked important information such as medication histories [35], environmental factors [25,30], and social determinants of health [28], which are known to play significant roles in health outcomes. Data-related limitations were also prevalent, with some studies dealing with the drawbacks of structured EHR data [4,26,33,35], potential of data misreporting [32], and missing data that could affect the reliability of their models [29,35]. In addition, from a clinical perspective, certain studies faced limitations owing to the lack of standardized definitions for specific outcomes [11,22,23,27,28], emphasizing the importance of consistent criteria in health care research such as in asthma management. The model explanation and implementation information are summarized in Table 5. All data extraction results can be found in Multimedia Appendix 1. We have also depicted some of the important principal findings in Multimedia Appendix 2.

Table 5. Summary of model explainability and implementation.

Study, year	Best model transparency	Model explanation methods	Follow reporting guidelines	Clinical implementation	Study limitations
Inselman, et al [27], 2022	Opaque model	SHAP ^a	No	No	<ul style="list-style-type: none"> Missing relevant variables Limited data about different biologics Diverse primary uses of biologics Heterogeneity in patient characteristics
Hurst et al [25], 2022	Opaque model	No model explanation	No	No	<ul style="list-style-type: none"> Missing relevant variables Single-center study Location-dependent model performance Limited environmental data
Hogan et al [28], 2022	Opaque model	Estimated weights	No	No	<ul style="list-style-type: none"> Missing relevant variables Lack of longitudinal outcomes Use of <i>ICD-9</i>^b (older clinical coding) Hospital differentiation Absence of demographic data and social determinants
Zein et al [29], 2021	Opaque model	SHAP	No	No	<ul style="list-style-type: none"> Limited generalizability Reliance on diagnostic codes Limited clinical information Exclusion of anti-IL5^c therapy Cross-sectional nature Quality of clinical information Limited PFT^d and FeNO^e data Handling missing data
Sills et al [30], 2021	Opaque model	autoML method	No	No	<ul style="list-style-type: none"> Retrospective nature Patient selection criteria Limited clinical information Exclusion of home and environmental factors Timing of posttriage variables
Hozawa et al [31], 2021	Opaque model	Extracted risk factors	No	No	<ul style="list-style-type: none"> Age distribution discrepancy Limitations of claim data Prevalent user design Causality estimation
Lisspers et al [32], 2021	Opaque model	LGBM ^f gain score	No	No	<ul style="list-style-type: none"> Data misreporting Applicability to other settings High false-positive rate Performance of shortlist model
Ananth et al [23], 2021	Transparent model	LR ^g coefficients	No	No	<ul style="list-style-type: none"> Lack of formal asthma control assessment Limited longitudinal outcomes Lack of comorbidity information
Tong et al [33], 2021	Opaque model	XGBoost ^h feature importance	No	No	<ul style="list-style-type: none"> Lack of relevant variables Nonuse of deep learning and unstructured data Expansion of data sources Generalizability across health care systems and diseases
Mehrish et al [24], 2021	Transparent model	No model explanation	No	No	<ul style="list-style-type: none"> Lack of relevant variables Limited method explanation

Study, year	Best model transparency	Model explanation methods	Follow reporting guidelines	Clinical implementation	Study limitations
Xiang et al [4], 2020	Opaque model	Attention mechanism	No	No	<ul style="list-style-type: none"> Absence of complex interactions among clinical variables Limitations of structured EHRⁱ data Challenges in distinguishing symptoms and risk factors Opportunities for model enhancement
Cobian et al [34], 2020	Transparent model	LR coefficients	No	No	<ul style="list-style-type: none"> Limited samples
Luo et al [35], 2020	Opaque model	XGBoost feature importance	No	No	<ul style="list-style-type: none"> Lack of medication claim data Limitations of structured EHR data Opportunities for additional features Data completeness and generalizability
Roe et al [22], 2020	Opaque model	XGBoost feature importance	No	No	<ul style="list-style-type: none"> Intensive care setting exclusivity Exclusion of routine intensive care features Generalizability to outpatient settings
Luo et al [26], 2020	Opaque model	XGBoost feature importance	No	No	<ul style="list-style-type: none"> Potential unexplored features Nonuse of deep learning and unstructured data Limited generalizability assessment
Wu et al [21], 2018	Opaque model	No model explanation	No	No	<ul style="list-style-type: none"> Suboptimal neural network configuration Limited scope Clinical relevance and feature weighting
Patel et al [11], 2018	Opaque model	GBDT ^j feature importance	No	No	<ul style="list-style-type: none"> Single institution data Pragmatic definition of the asthma population Lack of model validation Data limitations Lack of weather and CDC^k influenza data

^aSHAP: shapely additive explanation.

^bICD-9: *International Classification of Diseases, Ninth Revision*.

^cIL-5: interleukin 5.

^dPFT: Pulmonary Function Tests.

^eFeNO: Fractional Exhaled Nitric Oxide.

^fLGBM: light gradient boosting method.

^gLR: logistic regression.

^hXGBoost: extreme gradient boosting.

ⁱEHR: electronic health record.

^jGBDT: gradient boosting decision tree.

^kCDC: Centers for Disease Control and Prevention.

Discussion

Principal Findings

Our review indicates that this research field is still underdeveloped, given the limited body of evidence, heterogeneity of methods, lack of external validation, and suboptimally reported models. There was considerable heterogeneity in the specific definition of asthma outcome and the associated time horizon used by studies that sought to develop asthma attack risk prediction models. Class imbalance was also common across studies, and there was also considerable heterogeneity in how it was handled. Consequently, it was challenging to directly compare the studies.

The GBDT-based methods were the most reported best-performing method. DL methods such as long short-term memory (LSTM), a relatively more complex and advanced method, were also found in a few studies [4,21,34]. However, none of the studies compared the performance of the DL-based models with that of GBDT-based models. Moreover, none of the studies was prospectively evaluated or followed any reporting guidelines, and most studies (3/17, 18%) were not externally validated.

Strengths and Limitations

The key strengths of our study include undertaking a systematic and transparent approach to ensure reproducibility. Overall, 2 independent reviewers followed a clear framework during the study selection and data extraction stage. Furthermore, the interpretation of the result was supported by a multidisciplinary team consisting of both technical and clinical experts.

A further strength is that most systematic reviews about the use of ML methods in asthma research have focused on either diagnosis or classifying asthma subtypes [52-56]. Although there have been 2 previous reviews about the use of ML in predicting asthma attacks [15,16], our review is the first to focus on several key considerations in an ML pipeline, from data preprocessing to model implementation for asthma attack predictions.

However, this review also has 3 key limitations. First, this scoping review provided broad coverage of various technical challenges, but it cannot ascertain how feasible and effective an ML-based intervention can be in supporting asthma management. Second, we were not able to directly compare studies owing to the heterogeneity across studies, and that prohibited us from identifying the best algorithm or approach for solving the technical challenges highlighted in this review. Finally, this review only focused on the technical challenges without taking into account additional, crucial, sociocultural and organizational barriers to the adoption of ML-based tools in health care [57-59].

Interpretation in the Light of the Published Literature

The heterogeneity of outcome definitions found in this paper was also uncovered in previous non-ML asthma attack prognosis studies [16,60]. This heterogeneity includes both the indicators they used to define asthma attacks and the prediction time resolution. Recent systematic reviews also highlighted the wide

range of outcome variations in ML-based prognostic models for ischemic stroke [61] and brain injury [62].

GBDT methods, especially XGBoost, have become a state-of-the-art method, especially for large and structured data in several domains [63-65]. Among the DL methods, LSTM has also shown potential in several previous studies [66,67]. LSTM is one of the most popular methods for analyzing complex time series data. Its capability to learn the sequence pattern makes it very powerful to build a prediction model by representing the data as a sequence of events. EHR data consist of a sequence of historical clinical events, which represent the trajectory of each patient's condition over time. Incorporating the temporal features into the model, rather than just summarizing the events, can potentially boost the model's performance.

Most of the studies (14/17, 82%) in this review incorporated some form of model explainability that aimed to provide an accessible explanation of how the prediction is derived by the model to instill trust in the users [68]. Previous studies in various domains showed that an ML model can output a biased prediction caused by latent characteristics within the data [69]. Model explainability is therefore crucial to provide model transparency and enhance fairness [70], especially in high-stake tasks such as those in health care [71].

Model validation and standard reporting are some of the important challenges that can influence adoption into routine practices [72]. An ML model should be internally, externally, and prospectively validated to assess its robustness in predicting new data [73]. In addition, a standard guideline needs to be followed in reporting an ML model development [74] such as the Transparent Reporting of a Multivariable Prediction Model for Individual Prognosis or Diagnosis [75] or the Developmental and Exploratory Clinical Investigations of Decision Support Systems Driven by Artificial Intelligence [76]. It will facilitate an improved and objective understanding and interpretation of model performance. However, our review found a lack of external validation and adherence to reporting guidelines among the selected studies. These points resonated with the findings in other reviews of different cases [77,78].

Implications for Research, Policy, and Practice

This review highlighted several technical challenges that need to be addressed when developing asthma attack risk prediction algorithms. Further studies are required to develop a robust strategy for dealing with the class imbalance in asthma research. Class imbalance has been a common problem when working with EHR data [79,80]. However, there remains a notable gap in the literature regarding a systematic comparison of the effectiveness of existing methods, particularly in the context of asthma attack prediction. Several simple ML algorithms, such as linear regression, LR, and simple DTs, are easily interpretable [81]. In general, however, there is a trade-off between model interpretability and complexity, and most advanced methods are difficult to interpret, which then influences the users' perception and understanding of the model [82]. We believe that the black box nature of the more complex methods, such as XGBoost and LSTM, is likely a technical barrier to implementing such models in a real-world clinical setting.

Consequently, there is a need to continue exploring model explainability methods such as the attention mechanism approach recently developed for LSTM [83-85] that can augment complex “black box” algorithms.

There is a need for developing a global or at least a nationwide benchmark data set to facilitate external validation and to test the model’s generalizability [86]. Such validation is needed to ensure that the model will not only perform well under the data used in the model development but also can be reproduced to predict new data from different settings [87]. In addition, to maintain the transparency and reproducibility of the ML-based prediction model, adherence to a standard reporting guideline such as the Transparent Reporting of a Multivariable Prediction Model for Individual Prognosis or Diagnosis [75] should be encouraged. Both good reproducibility and clear reporting are key points to facilitate critical assessment of the model before its implementation into routine practices. This effort is pivotal in addressing ethical concerns associated with data-driven prediction tools and in guaranteeing the safety and impartiality of the prediction [88]. Ensuring the ethical aspects of integrating a data-driven model into routine clinical practice is becoming a great challenge. This task demands substantial resources and

relies on a collaborative effort involving experts from various disciplines [89].

Finally, to ensure that the ML-based model meets the requirements of the practices, a clear use case must be articulated. We found that almost all studies follow a clear clinical guideline to define asthma attacks, but there is a wide range of prediction time horizons across the studies. These variations are the result of distinct needs and goals from different practices. It is impossible to make a one-size-fits-all model. Therefore, a clear and specific clinical use case should be defined as the basis for developing an ML-based model.

Conclusions

ML model development for asthma attack prediction has been studied in recent years and includes the use of both traditional and DL methods. There is considerable heterogeneity in ML pipelines across existing studies that prohibits meaningful comparison. Our review indicates several key technical challenges that need to be tackled to make progress toward clinical implementation such as class imbalance problem, external validation, model explanation, and adherence to reporting guidelines for model reproducibility.

Acknowledgments

This paper presents independent research under the Asthma UK Centre for Applied Research (AUKCAR) funded by Asthma+Lung UK and Chief Scientist Office (CSO), Scotland (grant number: AUK-AC-2018-01). The views expressed are those of the authors and not necessarily those of Asthma+Lung UK or CSO, Scotland.

Conflicts of Interest

None declared.

Multimedia Appendix 1

List of the search keywords and full data extraction result.

[[XLSX File \(Microsoft Excel File\), 19 KB - ai_v2i1e46717_app1.xlsx](#)]

Multimedia Appendix 2

Key findings.

[[PDF File \(Adobe PDF File\), 84 KB - ai_v2i1e46717_app2.pdf](#)]

Multimedia Appendix 3

Preferred Reporting Items for Systematic Reviews and Meta-Analyses extension for Scoping Reviews (PRISMA-ScR) checklist.

[[PDF File \(Adobe PDF File\), 101 KB - ai_v2i1e46717_app3.pdf](#)]

References

1. Asthma. World Health Organization. 2023 May 04. URL: <https://www.who.int/news-room/fact-sheets/detail/asthma> [accessed 2023-11-28]
2. GBD 2019 Diseases and Injuries Collaborators. Global burden of 369 diseases and injuries in 204 countries and territories, 1990-2019: a systematic analysis for the Global Burden of Disease Study 2019. *Lancet* 2020 Oct 17;396(10258):1204-1222 [[FREE Full text](#)] [doi: [10.1016/S0140-6736\(20\)30925-9](https://doi.org/10.1016/S0140-6736(20)30925-9)] [Medline: [33069326](https://pubmed.ncbi.nlm.nih.gov/33069326/)]
3. Pocket guide for asthma management and prevention. Global Initiative for Asthma. URL: <https://ginasthma.org/pocket-guide-for-asthma-management-and-prevention/> [accessed 2023-11-20]
4. Xiang Y, Ji H, Zhou Y, Li F, Du J, Rasmy L, et al. Asthma exacerbation prediction and risk factor analysis based on a time-sensitive, attentive neural network: retrospective cohort study. *J Med Internet Res* 2020 Jul 31;22(7):e16981 [[FREE Full text](#)] [doi: [10.2196/16981](https://doi.org/10.2196/16981)] [Medline: [32735224](https://pubmed.ncbi.nlm.nih.gov/32735224/)]

5. Wark PA, Gibson PG. Asthma exacerbations. 3: pathogenesis. *Thorax* 2006 Oct;61(10):909-915 [FREE Full text] [doi: [10.1136/thx.2005.045187](https://doi.org/10.1136/thx.2005.045187)] [Medline: [17008482](https://pubmed.ncbi.nlm.nih.gov/17008482/)]
6. Martin MJ, Beasley R, Harrison TW. Towards a personalised treatment approach for asthma attacks. *Thorax* 2020 Dec;75(12):1119-1129. [doi: [10.1136/thoraxjnl-2020-214692](https://doi.org/10.1136/thoraxjnl-2020-214692)] [Medline: [32839286](https://pubmed.ncbi.nlm.nih.gov/32839286/)]
7. Noble M, Burden A, Stirling S, Clark AB, Musgrave S, Alsallakh MA, et al. Predicting asthma-related crisis events using routine electronic healthcare data: a quantitative database analysis study. *Br J Gen Pract* 2021 Nov 25;71(713):e948-e957 [FREE Full text] [doi: [10.3399/BJGP.2020.1042](https://doi.org/10.3399/BJGP.2020.1042)] [Medline: [34133316](https://pubmed.ncbi.nlm.nih.gov/34133316/)]
8. Tibble H, Tsanas A, Horne E, Horne R, Mizani M, Simpson CR, et al. Predicting asthma attacks in primary care: protocol for developing a machine learning-based prediction model. *BMJ Open* 2019 Jul 09;9(7):e028375 [FREE Full text] [doi: [10.1136/bmjopen-2018-028375](https://doi.org/10.1136/bmjopen-2018-028375)] [Medline: [31292179](https://pubmed.ncbi.nlm.nih.gov/31292179/)]
9. Hussain Z, Shah SA, Mukherjee M, Sheikh A. Predicting the risk of asthma attacks in children, adolescents and adults: protocol for a machine learning algorithm derived from a primary care-based retrospective cohort. *BMJ Open* 2020 Jul 23;10(7):e036099 [FREE Full text] [doi: [10.1136/bmjopen-2019-036099](https://doi.org/10.1136/bmjopen-2019-036099)] [Medline: [32709646](https://pubmed.ncbi.nlm.nih.gov/32709646/)]
10. Bose S, Kenyon CC, Masino AJ. Personalized prediction of early childhood asthma persistence: a machine learning approach. *PLoS One* 2021 Mar 1;16(3):e0247784 [FREE Full text] [doi: [10.1371/journal.pone.0247784](https://doi.org/10.1371/journal.pone.0247784)] [Medline: [33647071](https://pubmed.ncbi.nlm.nih.gov/33647071/)]
11. Patel SJ, Chamberlain DB, Chamberlain JM. A machine learning approach to predicting need for hospitalization for pediatric asthma exacerbation at the time of emergency department triage. *Acad Emerg Med* 2018 Dec;25(12):1463-1470 [FREE Full text] [doi: [10.1111/acem.13655](https://doi.org/10.1111/acem.13655)] [Medline: [30382605](https://pubmed.ncbi.nlm.nih.gov/30382605/)]
12. Shickel B, Tighe PJ, Bihorac A, Rashidi P. Deep EHR: a survey of recent advances in deep learning techniques for electronic health record (EHR) analysis. *IEEE J Biomed Health Inform* 2018 Sep;22(5):1589-1604 [FREE Full text] [doi: [10.1109/JBHI.2017.2767063](https://doi.org/10.1109/JBHI.2017.2767063)] [Medline: [29989977](https://pubmed.ncbi.nlm.nih.gov/29989977/)]
13. Giger ML. Machine learning in medical imaging. *J Am Coll Radiol* 2018 Mar;15(3 Pt B):512-520. [doi: [10.1016/j.jacr.2017.12.028](https://doi.org/10.1016/j.jacr.2017.12.028)] [Medline: [29398494](https://pubmed.ncbi.nlm.nih.gov/29398494/)]
14. Yu KH, Beam AL, Kohane IS. Artificial intelligence in healthcare. *Nat Biomed Eng* 2018 Oct;2(10):719-731. [doi: [10.1038/s41551-018-0305-z](https://doi.org/10.1038/s41551-018-0305-z)] [Medline: [31015651](https://pubmed.ncbi.nlm.nih.gov/31015651/)]
15. Alharbi ET, Nadeem F, Cherif A. Predictive models for personalized asthma attacks based on patient's biosignals and environmental factors: a systematic review. *BMC Med Inform Decis Mak* 2021 Dec 09;21(1):345 [FREE Full text] [doi: [10.1186/s12911-021-01704-6](https://doi.org/10.1186/s12911-021-01704-6)] [Medline: [34886852](https://pubmed.ncbi.nlm.nih.gov/34886852/)]
16. Bridge J, Blakey JD, Bonnett LJ. A systematic review of methodology used in the development of prediction models for future asthma exacerbation. *BMC Med Res Methodol* 2020 Feb 05;20(1):22 [FREE Full text] [doi: [10.1186/s12874-020-0913-7](https://doi.org/10.1186/s12874-020-0913-7)] [Medline: [32024484](https://pubmed.ncbi.nlm.nih.gov/32024484/)]
17. Duboue P. *The Art of Feature Engineering: Essentials for Machine Learning*. Cambridge, United Kingdom: Cambridge University Press; 2020.
18. Arksey H, O'Malley L. Scoping studies: towards a methodological framework. *Int J Soc Res Methodol* 2005;8(1):19-32. [doi: [10.1080/1364557032000119616](https://doi.org/10.1080/1364557032000119616)]
19. Eriksen MB, Frandsen TF. The impact of patient, intervention, comparison, outcome (PICO) as a search strategy tool on literature search quality: a systematic review. *J Med Libr Assoc* 2018 Oct;106(4):420-431 [FREE Full text] [doi: [10.5195/jmla.2018.345](https://doi.org/10.5195/jmla.2018.345)] [Medline: [30271283](https://pubmed.ncbi.nlm.nih.gov/30271283/)]
20. Leonardo R. PICO: model for clinical questions. *Evid Based Med* 2018;4(1):1-2. [doi: [10.4172/2471-9919.1000115](https://doi.org/10.4172/2471-9919.1000115)]
21. Wu S, Liu S, Sohn S, Moon S, Wi CI, Juhn Y, et al. Modeling asynchronous event sequences with RNNs. *J Biomed Inform* 2018 Jul;83:167-177 [FREE Full text] [doi: [10.1016/j.jbi.2018.05.016](https://doi.org/10.1016/j.jbi.2018.05.016)] [Medline: [29883623](https://pubmed.ncbi.nlm.nih.gov/29883623/)]
22. Roe KD, Jawa V, Zhang X, Chute CG, Epstein JA, Matelsky J, et al. Feature engineering with clinical expert knowledge: a case study assessment of machine learning model complexity and performance. *PLoS One* 2020 Apr 23;15(4):e0231300 [FREE Full text] [doi: [10.1371/journal.pone.0231300](https://doi.org/10.1371/journal.pone.0231300)] [Medline: [32324754](https://pubmed.ncbi.nlm.nih.gov/32324754/)]
23. Ananth S, Navarra A, Vancheeswaran R. S1 obese, non-eosinophilic asthma: frequent exacerbators in a real-world setting. *Thorax* 2021;76:A5-A6. [doi: [10.1136/thorax-2021-BTSAbstracts.7](https://doi.org/10.1136/thorax-2021-BTSAbstracts.7)]
24. Mehrish D, Sairamesh J, Hasson L, Sharma M. Combining weather and pollution indicators with insurance claims for identifying and predicting asthma prevalence and hospitalizations. 2021 Presented at: 4th International Conference on Human Interaction and Emerging Technologies: Future Applications (IHIET – AI 2021); April 28-30, 2021; Strasbourg, France. [doi: [10.1007/978-3-030-74009-2_58](https://doi.org/10.1007/978-3-030-74009-2_58)]
25. Hurst JH, Zhao C, Hostetler HP, Ghiasi Gorveh M, Lang JE, Goldstein BA. Environmental and clinical data utility in pediatric asthma exacerbation risk prediction models. *BMC Med Inform Decis Mak* 2022 Apr 22;22(1):108 [FREE Full text] [doi: [10.1186/s12911-022-01847-0](https://doi.org/10.1186/s12911-022-01847-0)] [Medline: [35459216](https://pubmed.ncbi.nlm.nih.gov/35459216/)]
26. Luo G, Nau CL, Crawford WW, Schatz M, Zeiger RS, Rozema E, et al. Developing a predictive model for asthma-related hospital encounters in patients with asthma in a large, integrated health care system: secondary analysis. *JMIR Med Inform* 2020 Nov 09;8(11):e22689 [FREE Full text] [doi: [10.2196/22689](https://doi.org/10.2196/22689)] [Medline: [33164906](https://pubmed.ncbi.nlm.nih.gov/33164906/)]
27. Inselman JW, Jeffery MM, Maddux JT, Lam RW, Shah ND, Rank MA, et al. A prediction model for asthma exacerbations after stopping asthma biologics. *Ann Allergy Asthma Immunol* 2023 Mar;130(3):305-311. [doi: [10.1016/j.anai.2022.11.025](https://doi.org/10.1016/j.anai.2022.11.025)] [Medline: [36509405](https://pubmed.ncbi.nlm.nih.gov/36509405/)]

28. Hogan AH, Brimacombe M, Mosha M, Flores G. Comparing artificial intelligence and traditional methods to identify factors associated with pediatric asthma readmission. *Acad Pediatr* 2022;22(1):55-61. [doi: [10.1016/j.acap.2021.07.015](https://doi.org/10.1016/j.acap.2021.07.015)] [Medline: [34329757](https://pubmed.ncbi.nlm.nih.gov/34329757/)]
29. Zein JG, Wu CP, Attaway AH, Zhang P, Nazha A. Novel machine learning can predict acute asthma exacerbation. *Chest* 2021 May;159(5):1747-1757 [FREE Full text] [doi: [10.1016/j.chest.2020.12.051](https://doi.org/10.1016/j.chest.2020.12.051)] [Medline: [33440184](https://pubmed.ncbi.nlm.nih.gov/33440184/)]
30. Sills MR, Ozkaynak M, Jang H. Predicting hospitalization of pediatric asthma patients in emergency departments using machine learning. *Int J Med Inform* 2021 Jul;151:104468. [doi: [10.1016/j.ijmedinf.2021.104468](https://doi.org/10.1016/j.ijmedinf.2021.104468)] [Medline: [33940479](https://pubmed.ncbi.nlm.nih.gov/33940479/)]
31. Hozawa S, Maeda S, Kikuchi A, Koinuma M. Exploratory research on asthma exacerbation risk factors using the Japanese claims database and machine learning: a retrospective cohort study. *J Asthma* 2022 Jul;59(7):1328-1337. [doi: [10.1080/02770903.2021.1923740](https://doi.org/10.1080/02770903.2021.1923740)] [Medline: [33926352](https://pubmed.ncbi.nlm.nih.gov/33926352/)]
32. Lisspers K, Ställberg B, Larsson K, Janson C, Müller M, Łuczko M, et al. Developing a short-term prediction model for asthma exacerbations from Swedish primary care patients' data using machine learning - based on the ARCTIC study. *Respir Med* 2021;185:106483 [FREE Full text] [doi: [10.1016/j.rmed.2021.106483](https://doi.org/10.1016/j.rmed.2021.106483)] [Medline: [34077873](https://pubmed.ncbi.nlm.nih.gov/34077873/)]
33. Tong Y, Messinger AI, Wilcox AB, Mooney SD, Davidson GH, Suri P, et al. Forecasting future asthma hospital encounters of patients with asthma in an academic health care system: predictive model development and secondary analysis study. *J Med Internet Res* 2021 Apr 16;23(4):e22796 [FREE Full text] [doi: [10.2196/22796](https://doi.org/10.2196/22796)] [Medline: [33861206](https://pubmed.ncbi.nlm.nih.gov/33861206/)]
34. Cobian A, Abbott M, Sood A, Sverchkov Y, Hanrahan L, Guilbert T, et al. Modeling asthma exacerbations from electronic health records. *AMIA Jt Summits Transl Sci Proc* 2020 May 30;2020:98-107 [FREE Full text] [Medline: [32477628](https://pubmed.ncbi.nlm.nih.gov/32477628/)]
35. Luo G, He S, Stone BL, Nkoy FL, Johnson MD. Developing a model to predict hospital encounters for asthma in asthmatic patients: secondary analysis. *JMIR Med Inform* 2020 Jan 21;8(1):e16080 [FREE Full text] [doi: [10.2196/16080](https://doi.org/10.2196/16080)] [Medline: [31961332](https://pubmed.ncbi.nlm.nih.gov/31961332/)]
36. 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](https://doi.org/10.1038/sdata.2016.35)] [Medline: [27219127](https://pubmed.ncbi.nlm.nih.gov/27219127/)]
37. Sun Y, Wong AK, Kamel MS. Classification of imbalanced data: a review. *Int J Pattern Recognit Artif Intell* 2009;23(04):687-719. [doi: [10.1142/S0218001409007326](https://doi.org/10.1142/S0218001409007326)]
38. Chawla NV, Bowyer KW, Hall LO, Kegelmeyer WP. SMOTE: synthetic minority over-sampling technique. *J Artif Intell Res* 2002;16:321-357. [doi: [10.1613/JAIR.953](https://doi.org/10.1613/JAIR.953)]
39. Spelman VS, Porkodi R. A review on handling imbalanced data. 2018 Presented at: 2018 International Conference on Current Trends towards Converging Technologies (ICCTCT); March 1-3, 2018; Coimbatore, India. [doi: [10.1109/icctct.2018.8551020](https://doi.org/10.1109/icctct.2018.8551020)]
40. Mohammed R, Rawashdeh J, Abdullah M. Machine learning with oversampling and undersampling techniques: overview study and experimental results. 2020 Presented at: 11th International Conference on Information and Communication Systems (ICICS); April 7-9, 2020; Irbid, Jordan. [doi: [10.1109/icics49469.2020.239556](https://doi.org/10.1109/icics49469.2020.239556)]
41. Breerton RG, Lloyd GR. Support vector machines for classification and regression. *Analyst* 2010 Feb;135(2):230-267. [doi: [10.1039/b918972f](https://doi.org/10.1039/b918972f)] [Medline: [20098757](https://pubmed.ncbi.nlm.nih.gov/20098757/)]
42. Xu M, Tantisira KG, Wu A, Litonjua AA, Chu JH, Himes BE, et al. Genome wide association study to predict severe asthma exacerbations in children using random forests classifiers. *BMC Med Genet* 2011 Jun 30;12:90 [FREE Full text] [doi: [10.1186/1471-2350-12-90](https://doi.org/10.1186/1471-2350-12-90)] [Medline: [21718536](https://pubmed.ncbi.nlm.nih.gov/21718536/)]
43. Ren J, Lee SD, Chen X, Kao B, Cheng R, Cheung D. Naive Bayes classification of uncertain data. 2009 Presented at: 2009 Ninth IEEE International Conference on Data Mining; December 6-9, 2009; Miami Beach, FL. [doi: [10.1109/icdm.2009.90](https://doi.org/10.1109/icdm.2009.90)]
44. Song YY, Lu Y. Decision tree methods: applications for classification and prediction. *Shanghai Arch Psychiatry* 2015 Apr 25;27(2):130-135 [FREE Full text] [doi: [10.11919/j.issn.1002-0829.215044](https://doi.org/10.11919/j.issn.1002-0829.215044)] [Medline: [26120265](https://pubmed.ncbi.nlm.nih.gov/26120265/)]
45. Guo G, Wang H, Bell D, Bi Y, Greer K. KNN model-based approach in classification. 2003 Presented at: OTM Confederated International Conferences CoopIS, DOA, and ODBASE 2003; November 3-7, 2003; Sicily, Italy. [doi: [10.1007/978-3-540-39964-3_62](https://doi.org/10.1007/978-3-540-39964-3_62)]
46. Wu YC, Feng JW. Development and application of artificial neural network. *Wireless Pers Commun* 2017 Dec 30;102:1645-1656. [doi: [10.1007/s11277-017-5224-x](https://doi.org/10.1007/s11277-017-5224-x)]
47. Liang X, Jacobucci R. Regularized structural equation modeling to detect measurement bias: evaluation of lasso, adaptive lasso, and elastic net. *Struct Equ Modeling Multidiscip J* 2019 Dec 12;27(5):722-734 [FREE Full text] [doi: [10.1080/10705511.2019.1693273](https://doi.org/10.1080/10705511.2019.1693273)]
48. Holmes G, Donkin A, Witten IH. WEKA: a machine learning workbench. 1994 Presented at: ANZIIS '94 - Australian New Zealand Intelligent Information Systems Conference; November 29-December 2, 1994; Brisbane, Australia. [doi: [10.1109/anzis.1994.396988](https://doi.org/10.1109/anzis.1994.396988)]
49. He X, Zhao K, Chu X. AutoML: a survey of the state-of-the-art. *Knowl Based Syst* 2021 Jan 05;212:106622. [doi: [10.1016/j.knosys.2020.106622](https://doi.org/10.1016/j.knosys.2020.106622)]
50. Chen T, Guestrin C. XGBoost: a scalable tree boosting system. 2016 Presented at: 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining; August 13-17, 2016; San Francisco, CA. [doi: [10.1145/2939672.2939785](https://doi.org/10.1145/2939672.2939785)]

51. Barredo Arrieta A, Díaz-Rodríguez N, Del Ser J, Bennetot A, Tabik S, Barbado A, et al. Explainable artificial intelligence (XAI): concepts, taxonomies, opportunities and challenges toward responsible AI. *Inf Fusion* 2020 Jun;58:82-115. [doi: [10.1016/j.inffus.2019.12.012](https://doi.org/10.1016/j.inffus.2019.12.012)]
52. Daines L, McLean S, Buelo A, Lewis S, Sheikh A, Pinnock H. Systematic review of clinical prediction models to support the diagnosis of asthma in primary care. *NPJ Prim Care Respir Med* 2019 May 09;29(1):19 [FREE Full text] [doi: [10.1038/s41533-019-0132-z](https://doi.org/10.1038/s41533-019-0132-z)] [Medline: [31073125](https://pubmed.ncbi.nlm.nih.gov/31073125/)]
53. Loymans RJ, Debray TP, Honkoop PJ, Termeer EH, Snoeck-Stroband JB, Schermer TR, et al. Exacerbations in adults with asthma: a systematic review and external validation of prediction models. *J Allergy Clin Immunol Pract* 2018;6(6):1942-52.e15 [FREE Full text] [doi: [10.1016/j.jaip.2018.02.004](https://doi.org/10.1016/j.jaip.2018.02.004)] [Medline: [29454163](https://pubmed.ncbi.nlm.nih.gov/29454163/)]
54. Luo G, Nkoy FL, Stone BL, Schmick D, Johnson MD. A systematic review of predictive models for asthma development in children. *BMC Med Inform Decis Mak* 2015 Nov 28;15:99 [FREE Full text] [doi: [10.1186/s12911-015-0224-9](https://doi.org/10.1186/s12911-015-0224-9)] [Medline: [26615519](https://pubmed.ncbi.nlm.nih.gov/26615519/)]
55. Smit HA, Pinart M, Antó JM, Keil T, Bousquet J, Carlsen KH, et al. Childhood asthma prediction models: a systematic review. *Lancet Respir Med* 2015 Dec;3(12):973-984. [doi: [10.1016/S2213-2600\(15\)00428-2](https://doi.org/10.1016/S2213-2600(15)00428-2)] [Medline: [26597131](https://pubmed.ncbi.nlm.nih.gov/26597131/)]
56. Exarchos KP, Beltsiou M, Votti CA, Kostikas K. Artificial intelligence techniques in asthma: a systematic review and critical appraisal of the existing literature. *Eur Respir J* 2020 Sep 3;56(3):2000521 [FREE Full text] [doi: [10.1183/13993003.00521-2020](https://doi.org/10.1183/13993003.00521-2020)] [Medline: [32381498](https://pubmed.ncbi.nlm.nih.gov/32381498/)]
57. Watson J, Hutyra CA, Clancy SM, Chandiramani A, Bedoya A, Ilangovan K, et al. Overcoming barriers to the adoption and implementation of predictive modeling and machine learning in clinical care: what can we learn from US academic medical centers? *JAMIA Open* 2020 Jul;3(2):167-172. [doi: [10.1093/jamiaopen/ooz046](https://doi.org/10.1093/jamiaopen/ooz046)]
58. Morrison K. Artificial intelligence and the NHS: a qualitative exploration of the factors influencing adoption. *Future Healthc J* 2021 Sep 02;8(3):e648-e654. [doi: [10.7861/fhj.2020-0258](https://doi.org/10.7861/fhj.2020-0258)]
59. Pumplun L, Fecho M, Wahl N, Peters F, Buxmann P. Adoption of machine learning systems for medical diagnostics in clinics: qualitative interview study. *J Med Internet Res* 2021 Oct 15;23(10):e29301. [doi: [10.2196/29301](https://doi.org/10.2196/29301)]
60. Alharbi F, Atkins A, Stanier C. Understanding the determinants of cloud computing adoption in Saudi healthcare organisations. *Complex Intell Syst* 2016 Jul 13;2(3):155-171. [doi: [10.1007/s40747-016-0021-9](https://doi.org/10.1007/s40747-016-0021-9)]
61. Zeng M, Oakden-Rayner L, Bird A, Smith L, Wu Z, Scroop R, et al. Pre-thrombectomy prognostic prediction of large-vessel ischemic stroke using machine learning: a systematic review and meta-analysis. *Front Neurol* 2022 Sep 8;13:945813. [doi: [10.3389/fneur.2022.945813](https://doi.org/10.3389/fneur.2022.945813)]
62. Cerasa A, Tartarisco G, Bruschetta R, Ciancarelli I, Morone G, Calabrò RS, et al. Predicting outcome in patients with brain injury: differences between machine learning versus conventional statistics. *Biomedicines* 2022 Sep 13;10(9):2267. [doi: [10.3390/biomedicines10092267](https://doi.org/10.3390/biomedicines10092267)]
63. Stemerman R, Arguello J, Brice J, Krishnamurthy A, Houston M, Kitzmiller R. Identification of social determinants of health using multi-label classification of electronic health record clinical notes. *JAMIA Open* 2021 Feb 9;4(3):ooaa069. [doi: [10.1093/jamiaopen/ooaa069](https://doi.org/10.1093/jamiaopen/ooaa069)]
64. Pan P, Li Y, Xiao Y, Han B, Su L, Su M, et al. Prognostic assessment of COVID-19 in the intensive care unit by machine learning methods: model development and validation. *J Med Internet Res* 2020 Nov 11;22(11):e23128. [doi: [10.2196/23128](https://doi.org/10.2196/23128)]
65. Muro S, Ishida M, Horie Y, Takeuchi W, Nakagawa S, Ban H, et al. Machine learning methods for the diagnosis of chronic obstructive pulmonary disease in healthy subjects: retrospective observational cohort study. *JMIR Med Inform* 2021 Jul 6;9(7):e24796. [doi: [10.2196/24796](https://doi.org/10.2196/24796)] [Medline: [34255684](https://pubmed.ncbi.nlm.nih.gov/34255684/)]
66. Dong X, Deng J, Rashidian S, Abell-Hart K, Hou W, Rosenthal RN, et al. Identifying risk of opioid use disorder for patients taking opioid medications with deep learning. *J Am Med Inform Assoc* 2021 Jul 30;28(8):1683-1693 [FREE Full text] [doi: [10.1093/jamia/ocab043](https://doi.org/10.1093/jamia/ocab043)] [Medline: [33930132](https://pubmed.ncbi.nlm.nih.gov/33930132/)]
67. Zhao J, Feng Q, Wu P, Lupu RA, Wilke RA, Wells QS, et al. Learning from longitudinal data in electronic health record and genetic data to improve cardiovascular event prediction. *Sci Rep* 2019 Jan 24;9:717. [doi: [10.1038/s41598-018-36745-x](https://doi.org/10.1038/s41598-018-36745-x)]
68. Petch J, Di S, Nelson W. Opening the black box: the promise and limitations of explainable machine learning in cardiology. *Can J Cardiol* 2022 Feb;38(2):204-213. [doi: [10.1016/j.cjca.2021.09.004](https://doi.org/10.1016/j.cjca.2021.09.004)]
69. Rudin C, Radin J. Why are we using black box models in AI when we don't need to? A lesson from an explainable AI competition. *Harvard Data Sci Rev* 2019 Nov 01;1(2). [doi: [10.1162/99608f92.5a8a3a3d](https://doi.org/10.1162/99608f92.5a8a3a3d)]
70. Rasheed K, Qayyum A, Ghaly M, Al-Fuqaha A, Razi A, Qadir J. Explainable, trustworthy, and ethical machine learning for healthcare: a survey. *Comput Biol Med* 2022 Oct;149:106043. [doi: [10.1016/j.combiomed.2022.106043](https://doi.org/10.1016/j.combiomed.2022.106043)]
71. 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. [doi: [10.1038/s42256-019-0048-x](https://doi.org/10.1038/s42256-019-0048-x)]
72. Verma AA, Murray J, Greiner R, Cohen JP, Shojania KG, Ghassemi M, et al. Implementing machine learning in medicine. *Can Med Assoc J* 2021 Aug 29;193(34):E1351-E1357. [doi: [10.1503/cmaj.202434](https://doi.org/10.1503/cmaj.202434)]
73. Cabitza F, Campagner A, Soares F, García de Gadiana-Romualdo L, Challa F, Sulejmani A, et al. The importance of being external. methodological insights for the external validation of machine learning models in medicine. *Comput Methods Programs Biomed* 2021 Sep;208:106288. [doi: [10.1016/j.cmpb.2021.106288](https://doi.org/10.1016/j.cmpb.2021.106288)]

74. Stevens LM, Mortazavi BJ, Deo RC, Curtis L, Kao DP. Recommendations for reporting machine learning analyses in clinical research. *Circ Cardiovasc Qual Outcomes* 2020 Oct 14;13(10). [doi: [10.1161/circoutcomes.120.006556](https://doi.org/10.1161/circoutcomes.120.006556)]
75. Collins GS, Reitsma JB, Altman DG, Moons KG. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): the TRIPOD Statement. *BMC Med* 2015;13(1):1. [doi: [10.1186/s12916-014-0241-z](https://doi.org/10.1186/s12916-014-0241-z)]
76. Vasey B, Nagendran M, Campbell B, Clifton DA, Collins GS, Denaxas S, et al. Reporting guideline for the early stage clinical evaluation of decision support systems driven by artificial intelligence: DECIDE-AI. *BMJ* 2022 May 18;377:e070904 [FREE Full text] [doi: [10.1136/bmj-2022-070904](https://doi.org/10.1136/bmj-2022-070904)] [Medline: [35584845](https://pubmed.ncbi.nlm.nih.gov/35584845/)]
77. Wynants L, Van Calster B, Collins GS, Riley RD, Heinze G, Schuit E, et al. Prediction models for diagnosis and prognosis of covid-19: systematic review and critical appraisal. *BMJ* 2020 Apr 07;369:m1328. [doi: [10.1136/bmj.m1328](https://doi.org/10.1136/bmj.m1328)] [Medline: [32265220](https://pubmed.ncbi.nlm.nih.gov/32265220/)]
78. Tsang KC, Pinnock H, Wilson AM, Shah SA. Application of machine learning algorithms for asthma management with mHealth: a clinical review. *J Asthma Allergy* 2022 Jun; Volume 15:855-873. [doi: [10.2147/jaa.s285742](https://doi.org/10.2147/jaa.s285742)]
79. Santiso S, Casillas A, Pérez A. The class imbalance problem detecting adverse drug reactions in electronic health records. *Health Informatics J* 2018 Sep 19;25(4):1768-1778. [doi: [10.1177/1460458218799470](https://doi.org/10.1177/1460458218799470)]
80. Tasci E, Zhuge Y, Camphausen K, Krauze AV. Bias and class imbalance in oncologic data—towards inclusive and transferrable AI in large scale oncology data sets. *Cancers* 2022 Jun 12;14(12):2897. [doi: [10.3390/cancers14122897](https://doi.org/10.3390/cancers14122897)]
81. Molnar C, Casalicchio G, Bischl B. Interpretable machine learning – a brief history, state-of-the-art and challenges. 2020 Presented at: ECML PKDD 2020 Workshops; September 14-18, 2020; Ghent, Belgium. [doi: [10.1007/978-3-030-65965-3_28](https://doi.org/10.1007/978-3-030-65965-3_28)]
82. Lauritsen SM, Kristensen M, Olsen MV, Larsen MS, Lauritsen KM, Jørgensen MJ, et al. Explainable artificial intelligence model to predict acute critical illness from electronic health records. *Nat Commun* 2020 Jul 31;11(1):3852. [doi: [10.1038/s41467-020-17431-x](https://doi.org/10.1038/s41467-020-17431-x)]
83. Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, et al. Attention is all you need. arXiv Preprint posted online June 12, 2017. [FREE Full text] [doi: [10.5860/choice.189890](https://doi.org/10.5860/choice.189890)]
84. Shen L, Zheng J, Lee EH, Shpanskaya K, McKenna ES, Atluri MG, et al. Attention-guided deep learning for gestational age prediction using fetal brain MRI. *Sci Rep* 2022 Jan 26;12(1):1408. [doi: [10.1038/s41598-022-05468-5](https://doi.org/10.1038/s41598-022-05468-5)]
85. Nguyen-Duc T, Mulligan N, Mannu GS, Bettencourt-Silva JH. Deep EHR spotlight: a framework and mechanism to highlight events in electronic health records for explainable predictions. *AMIA Jt Summits Transl Sci Proc* 2021 May 17;2021:475-484. [Medline: [34457163](https://pubmed.ncbi.nlm.nih.gov/34457163/)]
86. Futoma J, Simons M, Panch T, Doshi-Velez F, Celi LA. The myth of generalisability in clinical research and machine learning in health care. *Lancet Digit Health* 2020 Sep;2(9):e489-e492 [FREE Full text] [doi: [10.1016/S2589-7500\(20\)30186-2](https://doi.org/10.1016/S2589-7500(20)30186-2)] [Medline: [32864600](https://pubmed.ncbi.nlm.nih.gov/32864600/)]
87. Bates DW, Auerbach A, Schulam P, Wright A, Saria S. Reporting and implementing interventions involving machine learning and artificial intelligence. *Ann Intern Med* 2020 Jun 02;172(11_Supplement):S137-S144. [doi: [10.7326/m19-0872](https://doi.org/10.7326/m19-0872)]
88. Angers Schmid A, Zhou J, Theuermann K, Chen F, Holzinger A. Fairness and explanation in AI-informed decision making. *Mach Learn Knowl Extr* 2022 Jun 16;4(2):556-579. [doi: [10.3390/make4020026](https://doi.org/10.3390/make4020026)]
89. Obafemi-Ajayi T, Perkins A, Nanduri B, Wunsch DCII, Foster JA, Peckham J. No-boundary thinking: a viable solution to ethical data-driven AI in precision medicine. *AI Ethics* 2021 Nov 29;2(4):635-643. [doi: [10.1007/s43681-021-00118-4](https://doi.org/10.1007/s43681-021-00118-4)]

Abbreviations

DL: deep learning

DT: decision tree

EHR: electronic health record

GBDT: gradient boosting decision tree

LR: logistic regression

LSTM: long short-term memory

ML: machine learning

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

SHAP: shapely additive explanation

XGBoost: extreme gradient boosting

Edited by K El Emam, B Malin; submitted 23.02.23; peer-reviewed by N Mungoli, H Musawir; comments to author 03.08.23; revised version received 28.09.23; accepted 09.10.23; published 07.12.23.

Please cite as:

Budiarto A, Tsang KCH, Wilson AM, Sheikh A, Shah SA

Machine Learning–Based Asthma Attack Prediction Models From Routinely Collected Electronic Health Records: Systematic Scoping Review

JMIR AI 2023;2:e46717

URL: <https://ai.jmir.org/2023/1/e46717>

doi: [10.2196/46717](https://doi.org/10.2196/46717)

PMID: [38875586](https://pubmed.ncbi.nlm.nih.gov/38875586/)

©Arif Budiarto, Kevin C H Tsang, Andrew M Wilson, Aziz Sheikh, Syed Ahmar Shah. Originally published in JMIR AI (<https://ai.jmir.org>), 07.12.2023. This is an open-access article distributed under the terms of the Creative Commons Attribution License (<https://creativecommons.org/licenses/by/4.0/>), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR AI, is properly cited. The complete bibliographic information, a link to the original publication on <https://www.ai.jmir.org/>, as well as this copyright and license information must be included.

Original Paper

Application of Artificial Intelligence to the Monitoring of Medication Adherence for Tuberculosis Treatment in Africa: Algorithm Development and Validation

Juliet Nabbuye Sekandi^{1,2}, MD, MSc, DrPH; Weili Shi³, MSc; Ronghang Zhu⁴, MSc; Patrick Kaggwa⁵, BSc; Ernest Mwebaze^{6,7}, BSc, MSc, PhD; Sheng Li³, PhD

¹Department of Epidemiology and Biostatistics, College of Public Health, University of Georgia, Athens, GA, United States

²Global Health Institute, College of Public Health, University of Georgia, Athens, GA, United States

³School of Data Science, University of Virginia, Charlottesville, VA, United States

⁴School of Computing, College of Engineering & Franklin College of Arts and Sciences, University of Georgia, Athens, GA, United States

⁵Department of Epidemiology and Biostatistics, School of Public Health, Makerere University, Kampala, Uganda

⁶Sunbird AI, Kampala, Uganda

⁷Artificial Intelligence Research Lab, College of Computing and Information Science, Makerere University, Kampala, Uganda

Corresponding Author:

Juliet Nabbuye Sekandi, MD, MSc, DrPH

Global Health Institute

College of Public Health

University of Georgia

100 Foster Road

Athens, GA, 30602

United States

Phone: 1 706 542 5257

Email: jsekandi@uga.edu

Abstract

Background: Artificial intelligence (AI) applications based on advanced deep learning methods in image recognition tasks can increase efficiency in the monitoring of medication adherence through automation. AI has sparsely been evaluated for the monitoring of medication adherence in clinical settings. However, AI has the potential to transform the way health care is delivered even in limited-resource settings such as Africa.

Objective: We aimed to pilot the development of a deep learning model for simple binary classification and confirmation of proper medication adherence to enhance efficiency in the use of video monitoring of patients in tuberculosis treatment.

Methods: We used a secondary data set of 861 video images of medication intake that were collected from consenting adult patients with tuberculosis in an institutional review board–approved study evaluating video-observed therapy in Uganda. The video images were processed through a series of steps to prepare them for use in a training model. First, we annotated videos using a specific protocol to eliminate those with poor quality. After the initial annotation step, 497 videos had sufficient quality for training the models. Among them, 405 were positive samples, whereas 92 were negative samples. With some preprocessing techniques, we obtained 160 frames with a size of 224×224 in each video. We used a deep learning framework that leveraged 4 convolutional neural networks models to extract visual features from the video frames and automatically perform binary classification of adherence or nonadherence. We evaluated the diagnostic properties of the different models using sensitivity, specificity, F_1 -score, and precision. The area under the curve (AUC) was used to assess the discriminative performance and the speed per video review as a metric for model efficiency. We conducted a 5-fold internal cross-validation to determine the diagnostic and discriminative performance of the models. We did not conduct external validation due to a lack of publicly available data sets with specific medication intake video frames.

Results: Diagnostic properties and discriminative performance from internal cross-validation were moderate to high in the binary classification tasks with 4 selected automated deep learning models. The sensitivity ranged from 92.8 to 95.8%, specificity from 43.5 to 55.4%, F_1 -score from 0.91 to 0.92, precision from 88% to 90.1%, and AUC from 0.78 to 0.85. The 3D ResNet model had the highest precision, AUC, and speed.

Conclusions: All 4 deep learning models showed comparable diagnostic properties and discriminative performance. The findings serve as a reasonable proof of concept to support the potential application of AI in the binary classification of video frames to predict medication adherence.

(JMIR AI 2023;2:e40167) doi:[10.2196/40167](https://doi.org/10.2196/40167)

KEYWORDS

artificial intelligence; deep learning; machine learning; medication adherence; digital technology; digital health; tuberculosis; video directly observed therapy; video therapy

Introduction

Tuberculosis (TB) is a leading cause of death worldwide, with an estimated 10.6 million new cases of the disease and 1.7 million patients dying in 2021 [1]. The global *End TB* strategy set goals to eliminate disease, deaths, and burden by 2030 [2], but these could be out of reach if critical gaps in diagnosis, treatment, and care are not addressed. Medication adherence, defined as the extent to which a person's behavior regarding medication corresponds with agreed recommendations from a health care provider, is one of the barriers to TB control [3]. It is estimated that 33% to 50% of patients who start treatment become nonadherent to their prescribed medication regimens [4,5]. Nonadherence is associated with the emergence of drug resistance, prolonged infectiousness, treatment failure, and death, especially in the context of TB and HIV coinfection [6,7]. The existing interventions to mitigate poor medication adherence have limited effectiveness for a variety of reasons [5]. In Africa, a high patient load coupled with a severe shortage of health workers hampers proper monitoring and support of patients on TB treatment [8]. Digital adherence technologies have rapidly emerged as tools for improving the delivery of care in a variety of health care settings [2,9]. In 2017, the World Health Organization endorsed the use of video-based directly observed therapy (VDOT) as a suitable alternative to directly observed therapy for monitoring TB treatment and published guidance on its implementation [10]. VDOT overcomes geographic barriers because it enables the health providers to view patients' medication intake activity remotely, especially in the hard-to-reach populations [11-13]. It also enhances autonomy since patients can choose when and where they take their TB medications [14-16]. The limitation with asynchronous VDOT is the repetitive manual task of reviewing videos and confirming daily adherence [17]. Moreover, such classification tasks are accomplished by following a prespecified protocol [18]. In the face of high patient workloads, repetitive manual tasks could lead to inaccurate assessment and human fatigue. High workload is a recognized occupational stressor that has implications for the quality of care and patient outcomes [19]. The automation of routine processes is a well-known solution to increase efficiency in daily workflows. Therefore, more advanced tools such as artificial intelligence (AI) can be integrated with digital adherence technologies to accelerate widespread adoption and impact [20,21].

AI applications have the potential to transform health care in several clinical practice areas, primarily medical imaging [22]. First, AI tools can increase productivity and the efficiency of care delivery by streamlining workflows in the health care

systems [23]. Second, AI can help improve the experience of health care workers, enabling them to spend more time in direct patient care and reducing stress-related burnout [19]. Third, AI can support the faster delivery of care, by enhancing clinical decision-making, helping health care systems manage population health more proactively, and allocating resources to where they can have the largest impact [24]. Modern computer vision techniques powered by deep learning convolutional neural networks (DCNNs) can be applied to medical imaging, medical videos, and clinical deployment [25]. Deep learning techniques that process raw data to perform classification or detection tasks can make digital adherence monitoring in TB control more effective and efficient. DCNNs are state-of-the-art machine learning algorithms that have the ability to learn from input data to recognize intricate activities and patterns [26]. These characteristics make DCNNs powerful tools for recognition, classification, and prediction. Moreover, the features discovered by the models are not predetermined by human experts but rather by the patterns they learn from input data [27,28]. This concept can be applied to patterns in the videos of medication intake. However, the development and implementation of deep learning methods in health care remain largely limited because of a lack of access to large, well-curated, and labeled data sets. Additionally, specific technical knowledge, skills, and expertise required to develop deep learning models are often uncommon among health care professionals [27]. The goal of our pilot was to conduct a proof of concept for the development of an AI system that can perform routine classification tasks applicable to medication adherence. We expect that this initial step will be the basis for further development and validation of AI tools that will be used across treatments in chronic diseases in a variety of clinical settings.

Methods

Study Design, Population, and Data Sources

In this pilot study, a multidisciplinary team consisting of a physician scientist with expertise in TB medication adherence; 2 computer scientists with expertise in machine learning, computer vision, and deep learning models; and 3 graduate students in computer science evaluated the technical feasibility of applying AI to analyze a raw data set of videos from patients with TB taking medications. We used a secondary data set of 861 self-recorded medication intake videos collected as part of a pilot VDOT study of 51 patients with TB. The pilot study was conducted in Uganda.

Ethical Approval

The study was approved by the Institutional Review Board Office of Research, University of Georgia (number PROJECT00002406) and the Makerere University Higher Degrees, Research and Ethics Committee in Uganda (number 756).

Patient Recruitment and Enrollment

A cohort of adult male and female patients aged 18-65 years with a confirmed diagnosis of TB attending public clinics in Kampala, Uganda, were enrolled in VDOT pilot studies from July 2018 to December 2020. The study evaluated the effectiveness of VDOT in monitoring adherence where daily medication intake videos were collected with the patients' written consent. Further details on the eligibility criteria and sociodemographic characteristics of the patients contributing to the video data sets are published elsewhere [16].

Process of Annotation and Labeling of Medication Videos

First, a team of 3 trained video annotators with a computer science background evaluated the videos in the primary medication intake data set to create a new medication intake video data set. Using a systematic iterative process of review and discussions, the research team developed a protocol for video annotation de novo, since no specific protocols existed for medication videos. The team included the 3 trained student annotators, a senior computer scientist, and a physician with expertise in medication adherence. The protocol was summarized into 3 basic rules that guided labeling videos as

positive—actual medication ingestion activity, *negative*—no medication intake activities, or *ambiguous*—if no pills were seen but there was a blurry image of a face, as described in Table 1. We used the de novo standardized protocol for labeling videos. To control the quality of the annotation, we only considered videos where there was complete agreement of the classification across the 3 annotators to create the final video data set for model training and evaluation. After the annotation process, out of 861 videos, we kept 497 videos, which consisted of 405 (47%) positive videos and 92 (10%) negative videos. The sex and class distribution of videos that were kept in the final data set was as follows: of the 405 positive videos from 51 patients, 248 (61.2%) were from 28 male patients and 157 (38.7%) videos were from 23 female patients. Only 36 patients produced 92 negative videos; 48 (52%) were from 19 male patients, and 44 (48%) were from 17 female patients. The average distribution was 8 positive videos and 2 negative videos per patient. The outcome of this process resulted in the medication intake video data set that was used as a training data set for the deep learning model. Second, we divided the data set into training and validation subsets to assess the performance of our deep learning framework and baselines on medication adherence recognition. Furthermore, we analyzed the influence of different deep learning architectures in our framework on medication adherence recognition, classification, and prediction. It is important to note that the video annotation process is only required to construct the data set for model training and evaluation of this study. Once the deep learning model is trained, we do not need manual annotations anymore for the new videos, when using the proposed methods in practice.

Table 1. The rules for video annotation, labeling, and outcome of the video data set.

Labels	Description	Videos (N=861), n (%)
Positive: actual medication ingestion activities=adherence	<ul style="list-style-type: none"> Videos show clear visibility of the face, pill, and water bottle Patient exhibits clear action of taking pills and drinking water Good illumination 	405 (47)
Negative: no actual medication ingestion activities=nonadherence	<ul style="list-style-type: none"> Face of patient seen No pills are detected Patient does not put the pills into his or her mouth or there is no action of drinking water Good illumination 	92 (10)
Excluded videos	— ^a	364 (42.3)
Ambiguous or uncertain videos	<ul style="list-style-type: none"> Pills not seen Blurry faces and hands 	157 (18.2)
Poor quality videos	<ul style="list-style-type: none"> Poor illumination Face of patient not seen 	152 (17.7)
Damaged videos	Not reviewed	55 (6)

^aNot applicable.

Preprocessing of the Annotated Medication Intake Videos

Before we used AI tools to analyze the medication adherence of the patients, some techniques were implemented to preprocess the videos. The video-preprocessing stage was divided into 3 parts. In the first part, each video was converted to the mp4

format since the mp4 format is more convenient to process than the original format of the raw videos. Next, we adopted FFmpeg, a leading multimedia framework, to extract the video frames from each video with the mp4 format. Nevertheless, not all the video frames were relevant to the medication adherence, and the number of the video frames for each video was quite different, which also posed a problem in our study. In the end,

we manually extracted the same number of key video frames that were the most relevant to medication adherence. These video frames constituted the final data set for our AI experiments.

Model Development: Deep Learning Framework

Our deep learning framework for recognizing medication intake activities consisted of 2 parts: first, convolutional neural networks (CNNs) were used to extract visual features from medication intake videos; and second, support vector machine (SVM) [29] was adopted as a classifier to generate prediction scores for videos as shown in Figure 1. In particular, inspired by the huge success of deep learning models in image and video analysis, we used 2D CNN and 3D CNN models to extract the high-dimensional, spatiotemporal features from input videos. These models were pretrained on large-scale, labeled image or video data sets. Then, the SVM, an effective classifier, was trained to classify the extracted high-dimensional features. Our framework consisted of DCNNs pretrained with external data sets: Inception-v4 [30]; 3D ResNet, designed for lower complexity structure with so-called skip residual connections [31]; 3D ResNext [32]; and Inflated 3D [33]. These DCNNs are extensively used by the computer science community for extracting features from images and videos [34]. Specifically, Inception-v4 is pretrained on the ImageNet data set [35]. 3D ResNet, 3D ResNext, and Inflated 3D are pretrained on the Kinetics data set [36,37]. Besides, the sizes of the feature vectors from each model are different. For instance, the length of the feature vector generated from Inception-v4 is 1536, whereas

the length of the feature vector is 2048 from 3D ResNet and 3D ResNext. The details of the feature length are illustrated in Table 2. In the training stage, we trained the SVM with features extracted by the pretrained DCNNs from the training data set. In the testing stage, our trained model, which consists of a DCNN and SVM, generated prediction scores for videos from the testing data set to recognize the medication adherence. The generated prediction score is a decimal number between 0 and 1, which can be interpreted as the probability that the video represents a patient correctly ingesting their medication.

These DCNN models are designed primarily to extract the feature from images, but they cannot deal with videos directly, due to the 3D structure of video data. To tackle this problem, various 3D CNN models have been developed, in which the 2D convolution operation is extended to 3D convolution operation. The 3D ResNet and 3D ResNext used in our study are built on the 2D CNN model ResNet [31] that introduces the idea of residual connections. Figure 2 illustrates the building blocks of the ResNet, 3D ResNet, and 3D ResNext. All 3 blocks consist of 3 convolution layers followed by batch normalization [32], rectified linear unit [33], and identity mapping [31]. The major difference is that the 2D convolution kernels (1×1 and 3×3) in ResNet are modified to 3D convolution kernels ($1 \times 1 \times 1$ and $3 \times 3 \times 3$) in 3D ResNet and 3D ResNext. Compared to 3D ResNet, 3D ResNext introduces the group convolutions in the second layer of the block, which divides the feature maps into small groups. In practice, 3D ResNet and 3D ResNext are typically composed of multiple layers [30,31].

Figure 1. Illustration of deep learning framework with feature extractor CNNs and classifier SVM. Different grey colors represent labeled videos, and black color denotes unlabeled videos. CNN: convolution neural network; SVM: support vector machine.

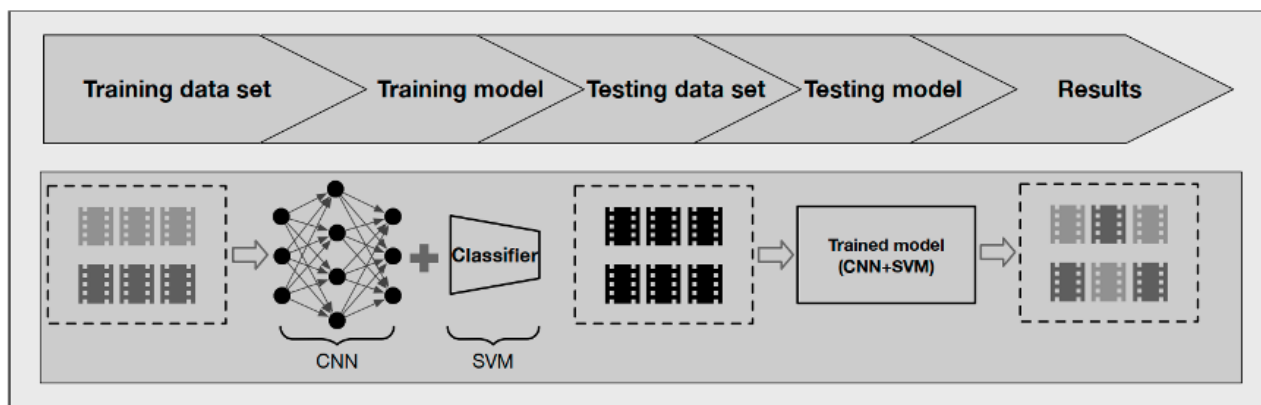
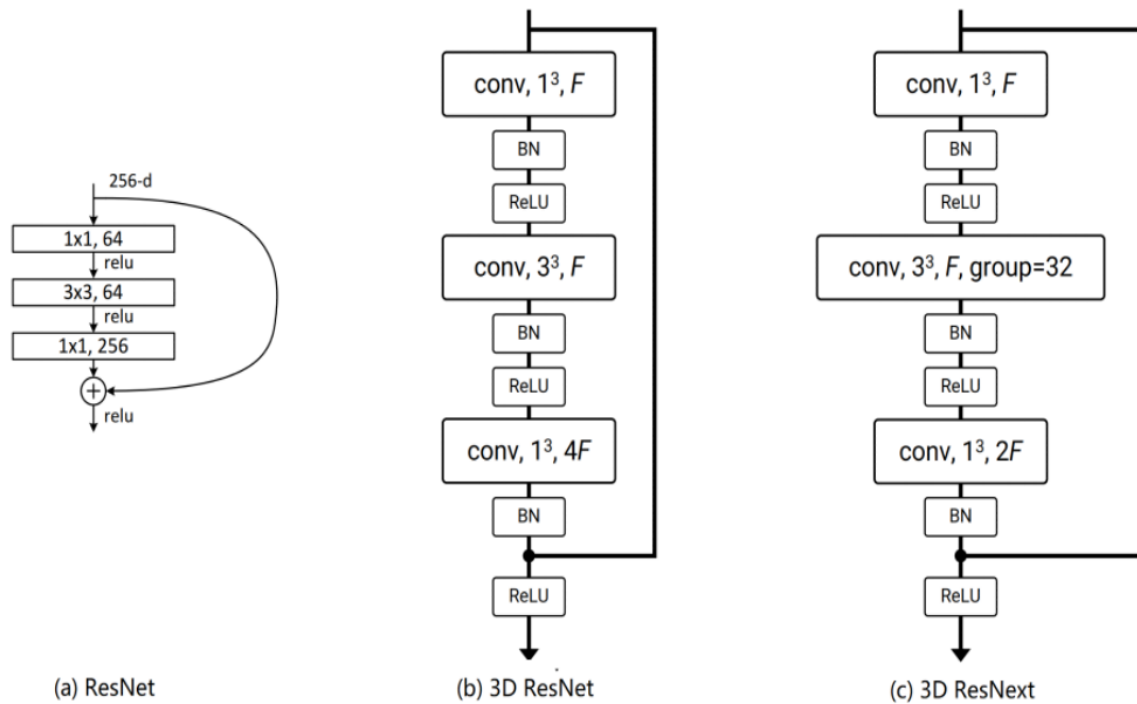


Table 2. The number of the features with its corresponding model.

Model	Features, n
HOG ^a	16,740
Inception-v4	1536
3D ResNet	2048
3D ResNext	2048
Inflated 3D	1024

^aHOG: histogram of oriented gradient.

Figure 2. Illustration of the building block of (a) ResNet, (b) 3D ResNet, and (c) 3D ResNext. BN: batch normalization; conv: convolution; F: number of feature channels; ReLU: rectified linear unit .



Apart from 3D ResNet and 3D ResNext, we also used Inception-v4 and Inflated 3D as our feature extractors. As a 2D CNN model, Inception-v4 is the fourth version of the Inception architecture network family. Compared to previous versions of the Inception family, Inception-v4 not only has a more uniformly simplified architecture and more inception modules but also absorbs the idea of residual connections from ResNet to form the new Inception block called residual inception blocks. Inflated 3D is another 3D CNN, which is built upon a 2D CNN from the Inception family. In our study, we compared the performance of one 2D CNN (Inception-v4) and three 3D CNNs (ie, 3D ResNet, 3D ResNext, and Inflated 3D). The 2D CNN treated each video as a set of video frames and generated a feature vector for each video frame, whereas 3D CNNs took video as a whole and generated a unified feature vector.

To better illustrate the effectiveness of deep learning models for medication adherence recognition, we used a traditional visual feature descriptor, histogram of oriented gradient (HOG) [38], as the replacement of the features extracted by DCNNs. HOG is a traditional descriptor that can generate handcrafted features directly from the images. The handcrafted feature was fed into the SVM for classification. In our pilot study, the SVM with HOG features was used as a baseline. Besides, we also investigated the average time of each method to extract features from the video frames, since efficiency is also an important indicator to evaluate the methods in practice.

Statistical Analysis

We adopted a 5-fold cross-validation strategy to evaluate the performance of our deep learning framework with different DCNNs as it is the recommended best practice for model validation [39]. We chose 5-fold cross-validation since it offers a good trade-off between efficiency and reliability, compared with alternative strategies such as leave-one-out cross-validation

or random splits. In the experiments, we evaluated the performance of our framework from different aspects by using 5 metrics: the area under the receiver operating characteristic (ROC) curve (AUC) and F_1 -score, which are primary evaluation metrics, and sensitivity (recall), specificity, and precision (positive predictive value), which are supplementary. The F_1 -score can be interpreted as the harmonic mean of precision and recall. We empirically set the threshold to 0.6 to neutralize the adverse effect of the imbalanced distribution of the data. For each given DCNN in our framework, we randomly split the data set into 5 subsets: 4 out of 5 subsets were used as the training data set, and the rest were adopted as the testing data set. We ran the 5-fold cross-validation 5 times. Each time, we randomly shuffled the order of the data before feeding the data into the model and reporting the mean values and SDs for each metric. Furthermore, another comparison experiment was implemented to show that our framework does not suffer from an overfitting problem with the high-dimensional features. Besides, we also drew the ROC curves to demonstrate the performance of different CNNs. We also evaluated the efficiency using speed in seconds as a metric defining the time required to extract features from the videos relevant to medications adherence. In addition, we noticed that metrics such as precision still have some limitations in the presence of class imbalance. This problem can be mitigated by adjusting the classification threshold.

Results

Performance in the Monitoring of Medication Adherence

3D ResNet achieved the best performance in the task of monitoring patient medication adherence activities as shown in

Table 3. The performance of 3D ResNext was very close to that of 3D ResNet since they both have similar structure. Besides, the results also reveal that 3D CNN models had better performance than the 2D CNN model and traditional feature descriptor method. Specifically, the HOG method obtained the lowest values on all metrics. It is noted that 3D ResNet, 3D ResNext, and Inflated 3D are specifically designed for video

feature extraction, whereas Inception-v4 is designed for image feature extraction. Overall, the performances of the 3D ResNet and 3D ResNext were very comparable in all the metrics. The 3D ResNet obtained the best results on the AUC, highlighting its advantage in the prediction of the medication adherence activity.

Table 3. Performance of the proposed deep learning framework under different convolution neural networks and histogram of oriented gradient (HOG).

Feature extractor	Sensitivity, mean (SD)	Specificity, mean (SD)	Precision, mean (SD)	F_1 -score, mean (SD)	AUC ^a , mean (SD)
HOG	90.77 (2.62)	27.35 (8.98)	85.03 (1.86)	87.77 (1.41)	0.65 (0.06)
Inception-v4	92.54 (3.53)	43.70 (8.64)	87.91 (1.95)	90.12 (1.90)	0.80 (0.05)
3D ResNet	<i>94.57^b</i> (2.61)	<i>54.57</i> (6.46)	<i>90.20</i> (1.81)	<i>92.30</i> (1.44)	<i>0.87</i> (0.04)
3D ResNext	94.17 (2.67)	51.74 (7.33)	89.62 (2.21)	91.81 (1.82)	0.85 (0.05)
Inflated 3D	92.94 (3.47)	49.78 (8.00)	89.08 (1.85)	90.94 (2.24)	0.82 (0.06)

^aAUC: area under the curve.

^bItalicized numbers represent the best result under each metric.

Assessing Overfitting of the Model

AI models usually suffer from the overfitting problem with high-dimensional features and limited number of training data. To further investigate whether high-dimensional features would cause the overfitting problem or not, we conducted additional experiments to give a better illustration. In this experiment, we used the pretrained 3D ResNet as the feature extractor and reduced the original feature dimension from 2048 to 256 with the principal component analysis method. The results are shown in [Table 4](#). We observed that both of dimensions achieved similar performance, which confirmed that our framework was not affected much by the overfitting problem.

The ROC curves in [Figure 3](#) were generated by plotting the true positive rate (sensitivity) against the false positive rate (specificity) at different threshold settings. The diagonal straight dashed line from (0,0) to (1,1) represents the performance of the random classifier. Ideally, all the ROC curves should lie above the straight dashed line. The further the curve deviates from the diagonal line, the better the classifier is. The curves in [Figure 3](#) can be divided into 3 groups. The first group representing 3D ResNet and 3D ResNext show that the 2 curves were the closest to the y-axis with the highest AUC. The second group consists of Inception-v4 and Inflated 3D, with AUCs of 0.78 and 0.80. The worst performing classifier was the traditional model HOG, which is very close to the diagonal line, and its AUC is only 0.60.

We also investigated the time efficiency of each method in our study and the results are illustrated in [Table 5](#). The machine

that ran the code consisted of 2 Intel E4208 CPUs and 1 P100 Tesla GPU. We evaluated the average time spent per video by each method to generate the relevant features. 3D ResNet was the fastest and took only 0.54 seconds to generate the features for each video, whereas HOG was the slowest, spending on average 4.53 seconds—8 times longer to generate the handcrafted features from a single video, signifying its inferiority in efficiency. The speeds of 3D ResNext and Inflated 3D were relatively comparable, whereas Inception-v4 was slower than the other DCNNs. Overall, considering both the model's accuracy and efficiency, 3D ResNet might be the better model because it has both high accuracy and efficiency of processing videos.

The class imbalance between positive and negative videos was pronounced in our data at a ratio of 405:92, respectively. To remedy the potential detrimental effect of the class imbalance in our data, we used a simple but effective method of adjusting the classification threshold [40]. We conducted experiments to illustrate how different threshold values affected the performance of our model. In the experiment, we used 3D ResNet as the feature extractor and chose 3 threshold values: 0.5, 0.6, and 0.7. Five-fold cross-validation with fixed splits was adopted as shown in [Table 6](#). We see that higher threshold values would lead to higher specificity and precision values but slightly lower sensitivity and F_1 -score values. Adjusting the classification threshold helped to balance the sensitivity and specificity.

Table 4. Performance of the proposed deep learning framework with different dimensions of features. 3D ResNet was adopted as the feature extractor.

Number of dimensions	Sensitivity	Specificity	Precision	F_1 -score	AUC ^a
256	93.09	51.09	89.39	91.12	0.83
2048	94.57	54.35	90.17	92.26	0.86

^aAUC: area under the curve.

Figure 3. Receiver operator curves for monitoring the medication adherence from models in our framework. AUC: area under the curve; HOG: histogram of oriented gradient.

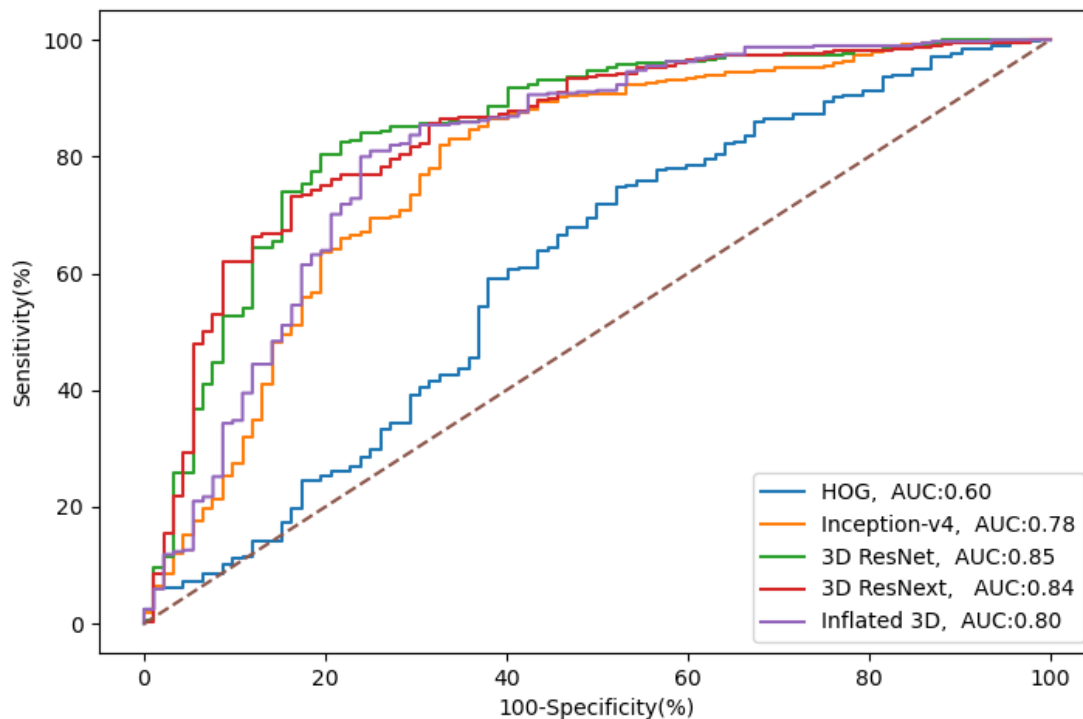


Table 5. The average time spent per video by each model.

Method	Time (seconds)
HOG ^a	4.53
Inception-v4	2.38
Inflated 3D	0.98
3D ResNext	0.6
3D ResNet	0.54

^aHOG: histogram of oriented gradient.

Table 6. Performance of the proposed deep learning framework with different classification thresholds. 3D ResNet was adopted as the feature extractor.

Threshold	Sensitivity	Specificity	Precision	F_1 -score
0.5	96.79	43.48	88.34	92.34
0.6	94.57	54.35	90.17	92.26
0.7	88.64	67.39	92.31	90.37

Discussion

Principal Finding

In this pilot project, we demonstrated a reasonable proof of concept that deep learning and AI techniques could be applied to advance support medication adherence monitoring. We tested 4 deep learning models and found that 3D ResNet performed best at an AUC of 0.84 and a speed of 0.54 seconds per video review. The level of discriminatory accuracy obtained is comparable to other machine learning algorithms that have been shown to achieve a diagnostic accuracy ranging from 72.5% to 77.3% in clinical settings. This level is similar to or higher than the expert clinical accuracy of doctors [41]. Spatiotemporal

models for action classification used in nonmedical fields have shown even better performance with an average accuracy of 90% [42]. A systematic review and meta-analysis of 69 studies comparing deep learning models against health care professionals concluded that both approaches were equivalent in diagnostic accuracy [43]. To our knowledge, this is the first pilot study to evaluate deep learning models for specific application to digital technologies and medication adherence in Africa.

Our model results could be limited by the relatively pronounced class imbalance between positive and negative samples in the data. To address the class imbalance problem, we adjusted the classification thresholds for the 3D ResNet model to better

balance the sensitivity and specificity. Specifically, we varied the thresholds at 0.5, 0.6, and 0.7 and found that across the range, sensitivity decreased slightly by 8% whereas specificity increased by 55%, thus improving the performance of the model. This means that by adjusting the classification threshold to 0.7, the model's ability to correctly identify persons who are not taking medications could be achieved. The relatively high performance of the deep learning models signifies the power of AI tools that can be harnessed for medication monitoring in routine clinical care or drug efficacy trials. We also acknowledge that our current experimental settings may lead to issues such as overfitting and data leakage, which are possible limitations to our findings. This could be due to the high dimensionality of features extracted by deep learning models and the small set of patients used in our study. In addition, the stratification is performed at the video level, and thus, it is possible that the videos from the same patient may appear in both training and test phases during cross-validation. Ideally, there is need to perform evaluations with stratification at the patient level; this step will be a priority in our future work. This pilot study is a valuable initial step for building more robust models that have relevant applications suitable for the local African context where the medication intake videos were collected. In the era of COVID-19 pandemic, the use of synchronous telehealth visits proved to be an extremely valuable care delivery approach when in-person provider-patient interactions were not possible [44,45]. Our proof-of-concept study explores the use of AI to bolster the utility of asynchronous remote provider-provider interactions. The evolving capacity of digital technologies to store and analyze various types of data will continue to revolutionize health care delivery in both resource-limited and resource-rich countries.

There are some strengths of this pilot study. For example, this is the first study that attempted to build and evaluate deep learning models using video images of TB medication intake from Uganda and the rest of Africa. We also developed a preliminary protocol for the annotation of medication video that can be refined further for use in low-income countries. This protocol was generated through a systematic iterative process of reviewing, discussing, and refining among a team of 3 trained video annotators who were computer science graduate students supervised by an expert in the field. Our pilot work builds on the existing literature and aspiration to expand the use of AI in routine health care [43] and, specifically, medication adherence monitoring [3]. By examining the utility of AI-based models, we are taking steps toward accelerating the future scale-up of digital adherence technologies in remote medication monitoring in TB, HIV/AIDS, and other chronic health conditions. The study was limited to the evaluation of the technical feasibility of developing a deep learning model. We did not incorporate all the recommended methodological features for the clinical validation of AI performance in real-world practice [46]. Indeed, we acknowledge that comprehensive validation is a critical next step for this work.

We also plan to develop new methods and evaluation protocols for the class-imbalanced settings in our future work.

It is worth noting that the same patient had multiple videos, which may introduce dependencies between images of the same patient and make the cross-validation less trustworthy. However, we clearly observed that the videos from the same patient had substantial differences in visual appearance. For example, some videos were recorded indoors whereas others were recorded outdoors, the same patient wore different clothes in different videos, and the viewpoints of video recording were also different. Furthermore, our method aimed to detect and understand the human medication adherence activities under a series of video frames. For instance, our model had to focus on specific key actions, for example, putting the pills into the mouth and drinking water, while trying to ignore the influence of the environment in the video frames. Although we used the video level to conduct the 5-fold cross-validation, the variance of the environment for videos from the same patient could present a challenge for our model to identify whether the patient has taken the pill or not.

Future Implications and Recommendations

Future work should be focused on improving the classification accuracy of deep learning models in medication adherence. First, there is a need for open-sourcing of large, labeled data sets with which to train the algorithms, especially in the African context. Second, additional techniques are needed to address class imbalance to improve the classification performance of deep learning models. Lastly, we propose to apply self-supervised learning methods, which provide a new way to pretrain DCNNs by exploiting pseudo-training labels that eliminates the time-consuming tasks of manual annotation. In our current deep learning framework, models are pretrained with external data sets, which may not be suitable for the extraction of visual features to classify medication adherence and nonadherence activities. All the neural network models showed comparable discriminative performance and diagnostic properties to state-of-the-art-performing deep learning algorithms. The findings serve as a reasonable proof of concept to support the potential utility of deep learning models in the binary classification of medication video frames to predict adherence. The success and widespread use of AI technologies will depend on data storage capacity, processing power, and other infrastructure capacities within health care systems [3]. Research is needed to evaluate the effectiveness of AI solutions in different patient groups and establish the barriers to widespread adoption of digital health technologies.

Conclusions

Our findings in this pilot study show the potential application of pretrained deep learning models and AI for the classification of medication adherence based on a unique video data set drawn in the African setting. The 3D ResNet model showed the best performance in relation to speed and discriminatory performance. Further development of AI tools to improve the monitoring of medication adherence could advance this field in public health, especially in low-resource settings.

Acknowledgments

We would like to thank Dr Esther Buregyeya, Dr Sarah Zalwango, and the field research team members in Uganda—Damalie Nakkonde, Gloria Nassanga, Daphine Kyaine, and Michelle Geno—for their assistance in collecting the video data for the research.

This work was supported by the National Center for Advancing Translational Sciences of the National Institutes of Health under award number UL1TR002378. The video data were collected with funding support from the National Institutes of Health Fogarty International Center under award number R21 TW011365. The funders had no role in the design, analysis, and interpretation of the study results. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

Authors' Contributions

JNS, WS, RZ, and SL researched literature and conceived the study. JNS was involved in seeking ethical approval and patient recruitment. JNS, WS, RZ, EM, SL, and PEK were involved in protocol development and data analysis. JNS and SL wrote the first draft of the manuscript. All authors reviewed and edited the manuscript and approved the final version of the manuscript.

Conflicts of Interest

None declared.

References

1. Global tuberculosis report 2022. World Health Organization. 2022 Oct 27. URL: <https://www.who.int/publications/i/item/9789240061729> [accessed 2023-02-07]
2. World Health Organization, European Respiratory Society. Digital health for the end TB strategy: an agenda for action. World Health Organization. 2015. URL: <https://apps.who.int/iris/handle/10665/205222> [accessed 2023-02-07]
3. Babel A, Taneja R, Mondello Malvestiti F, Monaco A, Donde S. Artificial intelligence solutions to increase medication adherence in patients with non-communicable diseases. *Front Digit Health* 2021 Jun 29;3:669869 [FREE Full text] [doi: [10.3389/fdgth.2021.669869](https://doi.org/10.3389/fdgth.2021.669869)] [Medline: [34713142](https://pubmed.ncbi.nlm.nih.gov/34713142/)]
4. Anuwatnonthakate A, Limsomboon P, Nateniyom S, Wattanaamornkiat W, Komsakorn S, Moolphate S, et al. Directly observed therapy and improved tuberculosis treatment outcomes in Thailand. *PLoS One* 2008 Aug 28;3(8):e3089 [FREE Full text] [doi: [10.1371/journal.pone.0003089](https://doi.org/10.1371/journal.pone.0003089)] [Medline: [18769479](https://pubmed.ncbi.nlm.nih.gov/18769479/)]
5. Alipanah N, Jarlsberg L, Miller C, Linh NN, Falzon D, Jaramillo E, et al. Adherence interventions and outcomes of tuberculosis treatment: a systematic review and meta-analysis of trials and observational studies. *PLoS Med* 2018 Jul 3;15(7):e1002595 [FREE Full text] [doi: [10.1371/journal.pmed.1002595](https://doi.org/10.1371/journal.pmed.1002595)] [Medline: [29969463](https://pubmed.ncbi.nlm.nih.gov/29969463/)]
6. Waitt CJ, Squire SB. A systematic review of risk factors for death in adults during and after tuberculosis treatment. *Int J Tuberc Lung Dis* 2011 Jul 01;15(7):871-885. [doi: [10.5588/ijtld.10.0352](https://doi.org/10.5588/ijtld.10.0352)] [Medline: [21496360](https://pubmed.ncbi.nlm.nih.gov/21496360/)]
7. Adane AA, Alene KA, Koye DN, Zeleke BM. Non-adherence to anti-tuberculosis treatment and determinant factors among patients with tuberculosis in northwest Ethiopia. *PLoS One* 2013 Nov 11;8(11):e78791 [FREE Full text] [doi: [10.1371/journal.pone.0078791](https://doi.org/10.1371/journal.pone.0078791)] [Medline: [24244364](https://pubmed.ncbi.nlm.nih.gov/24244364/)]
8. Bulage L, Sekandi J, Kigenyi O, Mupere E. The quality of tuberculosis services in health care centres in a rural district in Uganda: the providers' and clients' perspective. *Tuberc Res Treat* 2014 Sep 7;2014:685982-685911 [FREE Full text] [doi: [10.1155/2014/685982](https://doi.org/10.1155/2014/685982)] [Medline: [25276424](https://pubmed.ncbi.nlm.nih.gov/25276424/)]
9. WHO Global Observatory for eHealth. mHealth: new horizons for health through mobile technologies: second global survey on eHealth. World Health Organization. 2011. URL: <https://apps.who.int/iris/handle/10665/44607> [accessed 2023-02-07]
10. Guidelines for treatment of drug-susceptible tuberculosis and patient care (2017 update). World Health Organization. 2017. URL: <https://apps.who.int/iris/bitstream/handle/10665/255052/9789241550000-eng.pdf> [accessed 2023-02-07]
11. Garfein RS, Doshi RP. Synchronous and asynchronous video observed therapy (VOT) for tuberculosis treatment adherence monitoring and support. *J Clin Tuberc Other Mycobact Dis* 2019 Dec;17:100098 [FREE Full text] [doi: [10.1016/j.jctube.2019.100098](https://doi.org/10.1016/j.jctube.2019.100098)] [Medline: [31867442](https://pubmed.ncbi.nlm.nih.gov/31867442/)]
12. Story A, Aldridge RW, Smith CM, Garber E, Hall J, Ferenando G, et al. Smartphone-enabled video-observed versus directly observed treatment for tuberculosis: a multicentre, analyst-blinded, randomised, controlled superiority trial. *Lancet* 2019 Mar 23;393(10177):1216-1224 [FREE Full text] [doi: [10.1016/S0140-6736\(18\)32993-3](https://doi.org/10.1016/S0140-6736(18)32993-3)] [Medline: [30799062](https://pubmed.ncbi.nlm.nih.gov/30799062/)]
13. Story A, Garfein RS, Hayward A, Rusovich V, Dadu A, Soltan V, et al. Monitoring therapy compliance of tuberculosis patients by using video-enabled electronic devices. *Emerg Infect Dis* 2016 Mar;22(3):538-540 [FREE Full text] [doi: [10.3201/eid2203.151620](https://doi.org/10.3201/eid2203.151620)] [Medline: [26891363](https://pubmed.ncbi.nlm.nih.gov/26891363/)]
14. Garfein RS, Liu L, Cuevas-Mota J, Collins K, Muñoz F, Catanzaro DG, et al. Tuberculosis treatment monitoring by video directly observed therapy in 5 health districts, California, USA. *Emerg Infect Dis* 2018 Oct;24(10):1806-1815 [FREE Full text] [doi: [10.3201/eid2410.180459](https://doi.org/10.3201/eid2410.180459)] [Medline: [30226154](https://pubmed.ncbi.nlm.nih.gov/30226154/)]

15. Sinkou H, Hurevich H, Rusovich V, Zhylevich L, Falzon D, de Colombani P, et al. Video-observed treatment for tuberculosis patients in Belarus: findings from the first programmatic experience. *Eur Respir J* 2017 Mar 22;49(3):1602049 [FREE Full text] [doi: [10.1183/13993003.02049-2016](https://doi.org/10.1183/13993003.02049-2016)] [Medline: [28331042](https://pubmed.ncbi.nlm.nih.gov/28331042/)]
16. Sekandi JN, Buregyeya E, Zalwango S, Dobbin KK, Atuyambe L, Nakkonde D, et al. Video directly observed therapy for supporting and monitoring adherence to tuberculosis treatment in Uganda: a pilot cohort study. *ERJ Open Res* 2020 Jan 06;6(1):00175-2019 [FREE Full text] [doi: [10.1183/23120541.00175-2019](https://doi.org/10.1183/23120541.00175-2019)] [Medline: [32280670](https://pubmed.ncbi.nlm.nih.gov/32280670/)]
17. Garfein RS, Liu L, Cuevas-Mota J, Collins K, Catanzaro DG, Muñoz F, et al. Evaluation of recorded video-observed therapy for anti-tuberculosis treatment. *Int J Tuberc Lung Dis* 2020 May 01;24(5):520-525. [doi: [10.5588/ijtld.19.0456](https://doi.org/10.5588/ijtld.19.0456)] [Medline: [32398202](https://pubmed.ncbi.nlm.nih.gov/32398202/)]
18. National Center for HIV/AIDS, Viral Hepatitis, STD, and TB Prevention, Division of Tuberculosis Elimination. Implementing an electronic directly observed therapy (eDOT) program: a toolkit for tuberculosis programs. Centers for Disease Control and Prevention. 2015. URL: <https://www.cdc.gov/tb/publications/pdf/tbedottoolkit.pdf> [accessed 2023-02-07]
19. Erickson SM, Rockwern B, Koltov M, McLean RM, Medical PracticeQuality Committee of the American College of Physicians. Putting patients first by reducing administrative tasks in health care: a position paper of the American College of Physicians. *Ann Intern Med* 2017 May 02;166(9):659-661 [FREE Full text] [doi: [10.7326/M16-2697](https://doi.org/10.7326/M16-2697)] [Medline: [28346948](https://pubmed.ncbi.nlm.nih.gov/28346948/)]
20. Doshi R, Falzon D, Thomas BV, Temesgen Z, Sadasivan L, Migliori GB, et al. Tuberculosis control, and the where and why of artificial intelligence. *ERJ Open Res* 2017 Apr 21;3(2):00056-2017 [FREE Full text] [doi: [10.1183/23120541.00056-2017](https://doi.org/10.1183/23120541.00056-2017)] [Medline: [28656130](https://pubmed.ncbi.nlm.nih.gov/28656130/)]
21. Falzon D, Timimi H, Kurosinaki P, Migliori GB, Van Gemert W, Denkinger C, et al. Digital health for the End TB Strategy: developing priority products and making them work. *Eur Respir J* 2016 Jul;48(1):29-45 [FREE Full text] [doi: [10.1183/13993003.00424-2016](https://doi.org/10.1183/13993003.00424-2016)] [Medline: [27230443](https://pubmed.ncbi.nlm.nih.gov/27230443/)]
22. Davenport T, Kalakota R. The potential for artificial intelligence in healthcare. *Future Healthc J* 2019 Jun 13;6(2):94-98 [FREE Full text] [doi: [10.7861/futurehosp.6-2-94](https://doi.org/10.7861/futurehosp.6-2-94)] [Medline: [31363513](https://pubmed.ncbi.nlm.nih.gov/31363513/)]
23. 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](https://doi.org/10.1093/inthealth/ihaa007)] [Medline: [32300794](https://pubmed.ncbi.nlm.nih.gov/32300794/)]
24. Spatharou A, Hieronimus S, Jenkins J. Transforming healthcare with AI: the impact on the workforce and organizations. McKinsey & Company. 2020 Mar 10. URL: <https://www.mckinsey.com/industries/healthcare/our-insights/transforming-healthcare-with-ai> [accessed 2023-02-07]
25. Esteva A, Chou K, Yeung S, Naik N, Madani A, Mottaghi A, et al. Deep learning-enabled medical computer vision. *NPJ Digit Med* 2021 Jan 08;4(1):5 [FREE Full text] [doi: [10.1038/s41746-020-00376-2](https://doi.org/10.1038/s41746-020-00376-2)] [Medline: [33420381](https://pubmed.ncbi.nlm.nih.gov/33420381/)]
26. LeCun Y, Bengio Y, Hinton G. Deep learning. *Nature* 2015 May 28;521(7553):436-444. [doi: [10.1038/nature14539](https://doi.org/10.1038/nature14539)] [Medline: [26017442](https://pubmed.ncbi.nlm.nih.gov/26017442/)]
27. Faes L, Wagner SK, Fu DJ, Liu X, Korot E, Ledsam JR, et al. Automated deep learning design for medical image classification by health-care professionals with no coding experience: a feasibility study. *Lancet Digit Health* 2019 Sep;1(5):e232-e242 [FREE Full text] [doi: [10.1016/S2589-7500\(19\)30108-6](https://doi.org/10.1016/S2589-7500(19)30108-6)] [Medline: [33323271](https://pubmed.ncbi.nlm.nih.gov/33323271/)]
28. Mnih V, Kavukcuoglu K, Silver D, Rusu AA, Veness J, Bellemare MG, et al. Human-level control through deep reinforcement learning. *Nature* 2015 Feb 26;518(7540):529-533. [doi: [10.1038/nature14236](https://doi.org/10.1038/nature14236)] [Medline: [25719670](https://pubmed.ncbi.nlm.nih.gov/25719670/)]
29. Suthaharan S. Support vector machine. In: *Machine Learning Models and Algorithms for Big Data Classification*. Boston, MA: Springer; 2016:207-235.
30. Szegedy C, Ioffe S, Vanhoucke V, Alemi A. Inception-v4, Inception-ResNet and the impact of residual connections on learning. 2017 Feb 12 Presented at: Thirty-First AAAI Conference on Artificial Intelligence; February 4-9, 2017; San Francisco, CA. [doi: [10.1609/aaai.v31i1.11231](https://doi.org/10.1609/aaai.v31i1.11231)]
31. Hara K, Kataoka H, Satoh Y. Learning spatio-temporal features with 3D residual networks for action recognition. 2018 Jan 22 Presented at: 2017 IEEE International Conference on Computer Vision Workshops (ICCVW); October 22-29, 2017; Venice, Italy. [doi: [10.1109/iccvw.2017.373](https://doi.org/10.1109/iccvw.2017.373)]
32. Hara K, Kataoka H, Satoh Y. Can spatiotemporal 3D CNNs retrace the history of 2D CNNs and ImageNet? 2018 Dec 16 Presented at: 018 IEEE/CVF Conference on Computer Vision and Pattern Recognition; June 18-23, 2018; Salt Lake City, UT. [doi: [10.1109/cvpr.2018.00685](https://doi.org/10.1109/cvpr.2018.00685)]
33. Carreira J, Zisserman A. Quo vadis, action recognition? a new model and the kinetics dataset. 2017 Nov 9 Presented at: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR); July 21-26, 2017; Honolulu, HI. [doi: [10.1109/cvpr.2017.502](https://doi.org/10.1109/cvpr.2017.502)]
34. Bellemo V, Lim ZW, Lim G, Nguyen QD, Xie Y, Yip MYT, et al. Artificial intelligence using deep learning to screen for referable and vision-threatening diabetic retinopathy in Africa: a clinical validation study. *Lancet Digit Health* 2019 May;1(1):e35-e44 [FREE Full text] [doi: [10.1016/S2589-7500\(19\)30004-4](https://doi.org/10.1016/S2589-7500(19)30004-4)] [Medline: [33323239](https://pubmed.ncbi.nlm.nih.gov/33323239/)]
35. Jia DW, Dong W, Socher R, Li LJ, Li K, Li FF. ImageNet: a large-scale hierarchical image database. 2009 Aug 18 Presented at: 2009 IEEE Conference on Computer Vision and Pattern Recognition; June 20-25, 2009; Miami, FL. [doi: [10.1109/cvpr.2009.5206848](https://doi.org/10.1109/cvpr.2009.5206848)]
36. Kay W, Carreira J, Simonyan K, Zhang B, Hillier C, Vijayanarasimhan S, et al. The Kinetics Human Action Video Dataset. arXiv. Preprint posted online on May 19, 2017. [doi: [10.48550/arXiv.1705.06950](https://doi.org/10.48550/arXiv.1705.06950)]

37. He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition. 2016 Dec 12 Presented at: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR); June 27-30, 2016; Las Vegas, NV. [doi: [10.1109/cvpr.2016.90](https://doi.org/10.1109/cvpr.2016.90)]
38. Dalal N, Triggs B. Histograms of oriented gradients for human detection. 2005 Jul 25 Presented at: 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05); June 20-25, 2005; San Diego, CA. [doi: [10.1109/cvpr.2005.177](https://doi.org/10.1109/cvpr.2005.177)]
39. Hussain M. What is cross validation in machine learning? types of cross validation. Great Learning. 2020. URL: <https://www.mygreatlearning.com/blog/cross-validation/> [accessed 2023-02-07]
40. Johnson JM, Khoshgoftaar TM. Robust thresholding strategies for highly imbalanced and noisy data. 2022 Jan 25 Presented at: 2021 20th IEEE International Conference on Machine Learning and Applications (ICMLA); December 13-16, 2021; Pasadena, CA. [doi: [10.1109/icmla52953.2021.00192](https://doi.org/10.1109/icmla52953.2021.00192)]
41. Richens JG, Lee CM, Johri S. Improving the accuracy of medical diagnosis with causal machine learning. *Nat Commun* 2020 Aug 11;11(1):3923 [FREE Full text] [doi: [10.1038/s41467-020-17419-7](https://doi.org/10.1038/s41467-020-17419-7)] [Medline: [32782264](https://pubmed.ncbi.nlm.nih.gov/32782264/)]
42. Diba A, Fayyaz M, Sharma V, Arzani MM, Yousefzadeh R, Gall J, et al. Spatio-temporal channel correlation networks for action classification. 2018 Oct 6 Presented at: ECCV 2018: Computer Vision – ECCV 2018; September 8-14, 2018; Munich, Germany p. 299-315. [doi: [10.1007/978-3-030-01225-0_18](https://doi.org/10.1007/978-3-030-01225-0_18)]
43. Liu X, Faes L, Kale AU, Wagner SK, Fu DJ, Bruynseels A, et al. A comparison of deep learning performance against health-care professionals in detecting diseases from medical imaging: a systematic review and meta-analysis. *Lancet Digit Health* 2019 Oct;1(6):e271-e297 [FREE Full text] [doi: [10.1016/S2589-7500\(19\)30123-2](https://doi.org/10.1016/S2589-7500(19)30123-2)] [Medline: [33323251](https://pubmed.ncbi.nlm.nih.gov/33323251/)]
44. Kichloo A, Albosta M, Dettloff K, Wani F, El-Amir Z, Singh J, et al. Telemedicine, the current COVID-19 pandemic and the future: a narrative review and perspectives moving forward in the USA. *Fam Med Community Health* 2020 Aug 18;8(3):e000530 [FREE Full text] [doi: [10.1136/fmch-2020-000530](https://doi.org/10.1136/fmch-2020-000530)] [Medline: [32816942](https://pubmed.ncbi.nlm.nih.gov/32816942/)]
45. Doraiswamy S, Abraham A, Mamtani R, Cheema S. Use of telehealth during the COVID-19 pandemic: scoping review. *J Med Internet Res* 2020 Dec 01;22(12):e24087 [FREE Full text] [doi: [10.2196/24087](https://doi.org/10.2196/24087)] [Medline: [33147166](https://pubmed.ncbi.nlm.nih.gov/33147166/)]
46. Gulshan V, Peng L, Coram M, Stumpe MC, Wu D, Narayanaswamy A, et al. Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs. *JAMA* 2016 Dec 13;316(22):2402-2410. [doi: [10.1001/jama.2016.17216](https://doi.org/10.1001/jama.2016.17216)] [Medline: [27898976](https://pubmed.ncbi.nlm.nih.gov/27898976/)]

Abbreviations

AI: artificial intelligence
AUC: area under the curve
CNN: convolutional neural network
DCNN: deep learning convolutional neural network
HOG: histogram of oriented gradient
ROC: receiver operating characteristic
SVM: support vector machine
TB: tuberculosis
VDOT: video-based directly observed therapy

Edited by K El Emam; submitted 08.06.22; peer-reviewed by W Klement, Z Su; comments to author 19.07.22; revised version received 17.09.22; accepted 22.01.23; published 23.02.23.

Please cite as:

Sekandi JN, Shi W, Zhu R, Kaggwa P, Mwebaze E, Li S

Application of Artificial Intelligence to the Monitoring of Medication Adherence for Tuberculosis Treatment in Africa: Algorithm Development and Validation

JMIR AI 2023;2:e40167

URL: <https://ai.jmir.org/2023/1/e40167>

doi: [10.2196/40167](https://doi.org/10.2196/40167)

PMID: [38464947](https://pubmed.ncbi.nlm.nih.gov/38464947/)

©Juliet Nabbuye Sekandi, Weili Shi, Ronghang Zhu, Patrick Kaggwa, Ernest Mwebaze, Sheng Li. Originally published in JMIR AI (<https://ai.jmir.org>), 23.02.2023. This is an open-access article distributed under the terms of the Creative Commons Attribution License (<https://creativecommons.org/licenses/by/4.0/>), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR AI, is properly cited. The complete bibliographic information, a link to the original publication on <https://www.ai.jmir.org/>, as well as this copyright and license information must be included.

Original Paper

Effect of Benign Biopsy Findings on an Artificial Intelligence–Based Cancer Detector in Screening Mammography: Retrospective Case-Control Study

Athanasios Zouzos¹, MD; Aleksandra Milovanovic¹, MD; Karin Dembrower¹, MD, PhD; Fredrik Strand¹, MD, PhD

Department of Oncology and Pathology, Karolinska Institute, Stockholm, Sweden

Corresponding Author:

Athanasios Zouzos, MD

Department of Oncology and Pathology

Karolinska Institute

Solnavagen 1

Stockholm, 171 77

Sweden

Phone: 46 729142636

Email: athanasios.zouzos@ki.se

Abstract

Background: Artificial intelligence (AI)–based cancer detectors (CAD) for mammography are starting to be used for breast cancer screening in radiology departments. It is important to understand how AI CAD systems react to benign lesions, especially those that have been subjected to biopsy.

Objective: Our goal was to corroborate the hypothesis that women with previous benign biopsy and cytology assessments would subsequently present increased AI CAD abnormality scores even though they remained healthy.

Methods: This is a retrospective study applying a commercial AI CAD system (Insight MMG, version 1.1.4.3; Lunit Inc) to a cancer-enriched mammography screening data set of 10,889 women (median age 56, range 40–74 years). The AI CAD generated a continuous prediction score for tumor suspicion between 0.00 and 1.00, where 1.00 represented the highest level of suspicion. A binary read (flagged or not flagged) was defined on the basis of a predetermined cutoff threshold (0.40). The flagged median and proportion of AI scores were calculated for women who were healthy, those who had a benign biopsy finding, and those who were diagnosed with breast cancer. For women with a benign biopsy finding, the interval between mammography and the biopsy was used for stratification of AI scores. The effect of increasing age was examined using subgroup analysis and regression modeling.

Results: Of a total of 10,889 women, 234 had a benign biopsy finding before or after screening. The proportions of flagged healthy women were 3.5%, 11%, and 84% for healthy women without a benign biopsy finding, those with a benign biopsy finding, and women with breast cancer, respectively ($P<.001$). For the 8307 women with complete information, radiologist 1, radiologist 2, and the AI CAD system flagged 8.5%, 6.8%, and 8.5% of examinations of women who had a prior benign biopsy finding. The AI score correlated only with increasing age of the women in the cancer group ($P=.01$).

Conclusions: Compared to healthy women without a biopsy, the examined AI CAD system flagged a much larger proportion of women who had or would have a benign biopsy finding based on a radiologist's decision. However, the flagging rate was not higher than that for radiologists. Further research should be focused on training the AI CAD system taking prior biopsy information into account.

(*JMIR AI* 2023;2:e48123) doi:[10.2196/48123](https://doi.org/10.2196/48123)

KEYWORDS

artificial intelligence; AI; mammography; breast cancer; benign biopsy; screening; cancer screening; diagnostic; radiology; detection system

Introduction

Breast cancer is the most common cancer among women worldwide. It ranks fifth as a cause of cancer deaths because of its relatively favorable prognosis, but in the last 20 years, the average annual increase in breast cancer incidence rate has been 1.4% [1-3]. Screening programs have been clearly proven to reduce the mortality rate for breast cancer [4-6]. Retrospective studies have shown that outcomes might improve when radiologists combine mammography readings with an artificial intelligence (AI) system for computer-aided detection (CAD) [7-9]. Furthermore, reducing reading time with the assistance of an AI CAD system is possible [10,11]. An AI CAD system can be highly accurate for reading mammograms, and some systems are now on a comparable level with average breast radiologists at detecting breast cancer on screening mammography [12].

In addition to the well-known risk factors of age, family history, and hormonal history, there are also studies showing that benign breast disease increases the risk of breast cancer [13,14]. A study that analyzed risk factors for breast cancer found that having undergone any prior breast procedure was associated with an increased risk of breast cancer [15]. Another study showed that women found to have false-positive mammography findings were more likely to develop interval cancer or cancer at the second screening compared to those not recalled [16].

Radiologists performing screen reading normally have access to information about prior biopsies, while AI CAD systems do not take this information into account. In this retrospective study, we analyzed primarily to what extent the malignancy assessments of an AI CAD system are affected by the presence or absence of biopsy-proven benign findings. In a secondary analysis, we determined whether this effect differs between an AI CAD system and radiologists.

Methods

Study Population

This retrospective study was based on a case-control subset from the Cohort of Screen-Aged Women (CSAW). The CSAW is a complete population-based cohort of women aged 40 to 74 years invited to screening in the Stockholm region, Sweden, between 2008 and 2015 [17]. The exclusion criteria in the CSAW were having a prior history of breast cancer, having a diagnosis outside the screening range, and having had incomplete mammographic examinations. From the CSAW, a case-control subset was separately defined to contain all women from Karolinska University Hospital, Stockholm, who were diagnosed with breast cancer (n=1303), those at screening or clinical evaluation during the interval before the next planned screening, and 10,000 randomly selected healthy controls [17]. The purpose of the case-control subset is to make evaluation more efficient by not having to process an unnecessary amount of healthy controls while preserving the representability of the CSAW screening cohort in which it is nested. Additional exclusion criteria for the current study were having implants and receiving a cancer diagnosis later than 12 months after

mammography. The study population was divided into 3 groups based on their status: cancer, benign biopsy, and normal.

The cancer group was defined as having biopsy-verified breast cancer at screening or within 12 months of screening. The most recent mammographic screening prior to diagnosis was selected for analysis. The benign biopsy group was defined as having had a benign biopsy finding without ever having had breast cancer. The group was further stratified by the interval between biopsy and mammography. The normal group had neither breast cancer nor a prior benign biopsy finding. Women in the screening program who were previously recalled and deemed as having benign disease were also included in this group.

Mammography Assessments

The screening system consisted of double-reading followed by consensus discussion for any flagged examination. The following screening decision data were collected: flagging of abnormal screening by one or both radiologists and the final recall decision after consensus discussion. Screening decisions and clinical outcome data were collected by linking to regional cancer center registers.

AI CAD system

The AI CAD system was an Insight MMG (version 1.1.4.3; Lunit Inc). The reason for choosing Insight MMG for this study was that it demonstrated superior results in a retrospective analysis published in 2020 [9], which compared 3 AI CAD systems with a sensitivity and specificity comparable to Breast Cancer Surveillance Consortium benchmarks [18]. Briefly, the AI CAD system was originally trained on 170,230 mammograms from 36,468 women diagnosed with breast cancer and 133,762 healthy controls. The AI CAD system had been validated by previous studies using a deep learning model to triage screening mammograms [11,19]. The mammograms in the original training set were sourced from 5 institutions: 3 from South Korea, 1 from the United States, and 1 from the United Kingdom. The mammograms were acquired on mammography equipment from GE Healthcare, Hologic, and Siemens, and there were both screening and diagnostic mammograms. The generated prediction score for tumor presence was a decimal number between 0.00 and 1.00, where 1.00 represented the highest level of suspicion. The program assessed 2 images of each breast, and the highest score among the 4 images was selected to represent the examination. To obtain a binary assessment, determining whether the examination should be considered flagged for further workup by the AI CAD system, a cutoff point is required, above which the examination is considered flagged and below which the examination is considered not flagged. The cutoff point (0.40; AI abnormality threshold) defined whether an examination was considered flagged or not flagged by an AI CAD system, and was predefined in a prior study [9]. The cutoff was selected to enforce that the specificity of the AI CAD system should be the same as that for the average radiologist in that study. The examinations in the prior study originated from the same institution and are partly overlapping, which should ensure that the cutoff value is transferrable to the current setting.

Data Collection

The Stockholm-Gotland Regional Cancer Center provided personal identification numbers for all women who fulfilled the inclusion criteria for the CSAW. The identification numbers were linked to the local breast cancer quality register, “Regional Cancercentrum Stockholm-Gotlands Kvalitetsregister för Bröstcancer,” to collect data about breast cancer diagnosis. All diagnoses of breast cancer were biopsy verified. Benign diagnoses were collected from hospital electronic health records. All images were 2D full-field digital mammograms acquired on Hologic mammography equipment. The personal identification numbers were also linked to the radiological image repository to extract all digital mammograms from the Picture Archiving and Communication System.

Statistical Analysis

Statistical analysis was performed per patient and not per lesion. Stata (version 14 or later; StataCorp) was used for statistical analyses. The Wilcoxon rank-sum test and quantile regression analysis were used to examine differences between groups. To perform statistical tests, differences in medians were chosen due to the skewed distribution of AI scores. The required level for statistical significance was not adjusted for multiple comparisons. A value of $P < .05$ was considered statistically significant.

Ethics Approval

The collection and use of the data set by AI was approved by the Swedish Ethical Review Board (2017-02-08), and the need for informed consent was waived (diary number 2016/2600-31).

Results

We evaluated 11,303 women for inclusion in this retrospective case-control study (Figure 1). Of them, 414 women were excluded. The exclusion criteria were no mammographic examination in conjunction with a cancer diagnosis, having implants, and having cancer more than 12 months after mammography. The cancer group consisted of a total of 917 women, the benign biopsy group comprised 234 women, and the group with no cancer or biopsy (control group) comprised 9738 women.

Of the remaining 10,889 women, 8269 had complete information regarding radiologist assessments (when performing data collection, we received radiologist assessments only until December 31, 2015), which included selections rendered as potentially pathological by 1 or both radiologists and a final recall decision after consensus discussion. From those 8269 women, there were 724 women in the cancer group, 212 in the benign biopsy group, and 7371 in the normal mammography group.

There was a significant difference ($P < .001$) in AI scores among the cancer, benign biopsy, and normal mammography groups (Table 1).

Figure 1. Study population with exclusion criteria and subgroups.

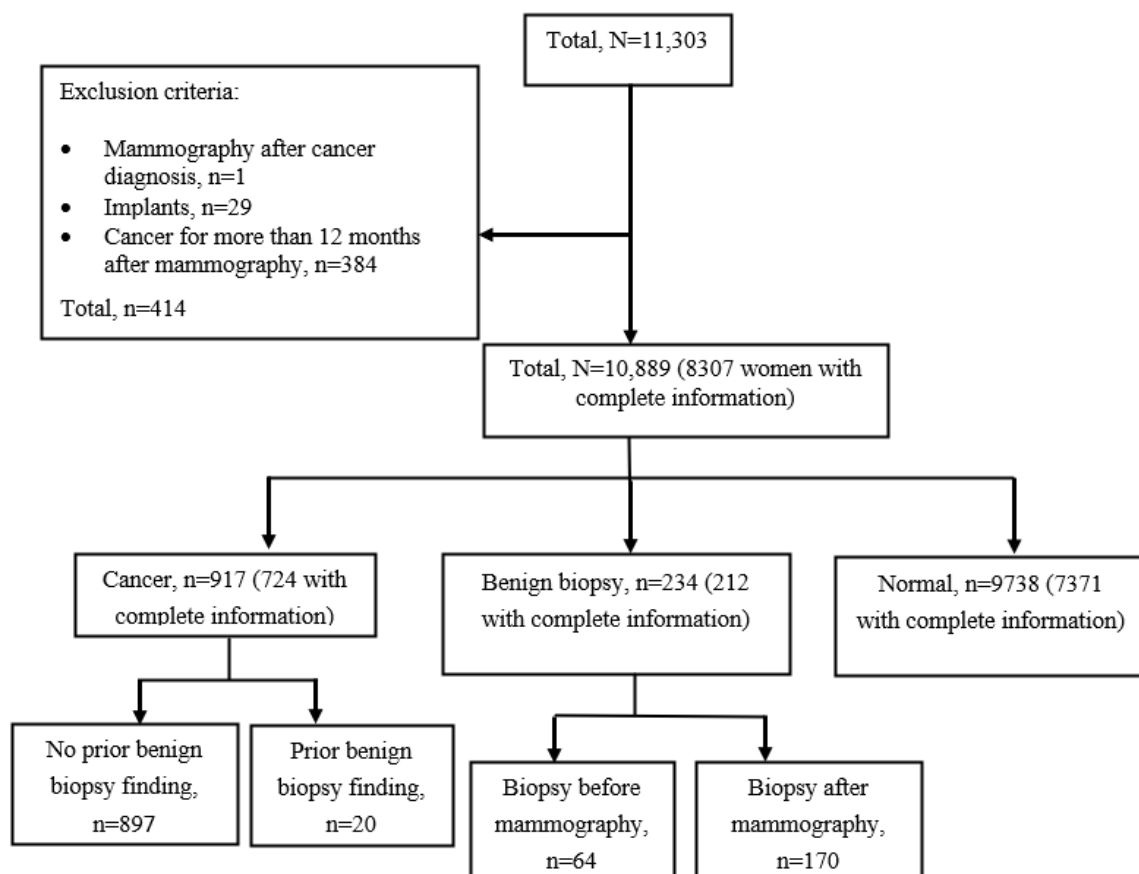


Table 1. Characteristics of the study population.

Characteristics	Participants, n	Age (years), median (IQR)	Proportion of assessments above the cut-off point, % (n/n)	AI ^a score, median (IQR)	P value (comparison with the cancer group)	P value (comparison with the normal biopsy group)
Normal with benign biopsy	234	51.2 (45.6-53.0)	11 (26/234)	0.051 (0.016-0.174)	<.001	<.001
Cancer	917	60.6 (50.7-66.6)	83 (768/917)	0.933 (0.666-0.983)	N/A ^b	<.001
Normal without biopsy	9738	55.5 (48.8-65.2)	3.5 (345/9738)	0.018 (0.005-0.065)	<.001	N/A

^aAI: artificial intelligence.

^bN/A: not applicable.

The proportion of AI assessments above the cutoff point was 3.5% in the group with normal mammography findings and 83% in the cancer group. In the benign biopsy group, 11% of the AI assessments were above the cutoff point. The distribution of AI scores for women diagnosed with breast cancer is shown in [Multimedia Appendix 1](#), that for healthy women with a benign biopsy in [Multimedia Appendix 2](#), and that for healthy women without a benign biopsy who remained healthy in [Multimedia Appendix 3](#).

In [Table 2](#), we show how the AI score is associated with the age of the women. There was a significant increase of the AI score in relation to age category in the cancer group ($P<.05$). There was no significant increase in the AI score in relation to age in the group with normal mammographic findings or in the group with benign biopsy findings. The median age for the study population was 56 years, and the median AI score was 0.023. The median age for the cancer group was 61 years, and the median AI score was 0.933 ($P=.01$). The median age of the group with previous benign biopsies was 49 years, and the median AI score was 0.051 ($P=.71$). The median AI score for healthy women was 0.018 ($P=.40$), and the median age of that group was 59 years.

The benign biopsy group was stratified by the interval between biopsy and mammography into 3 categories: 0-6 months, 6-24

months, and more than 24 months. There was no significant difference among the time-stratified categories ([Table 3](#)). In [Table 3](#), we describe the AI score related to the time between biopsy and mammography. Within 6 months after mammography, 104 of 234 participants had had a benign biopsy finding. The proportion of women with AI scores above the threshold was 16% for those with a benign biopsy finding within 6 months from mammography and 33% for those with a benign biopsy finding 6 months before mammography.

In the group with a benign biopsy finding after mammography, the proportion of abnormal assessments by AI, above cutoff point, was 15%, while the radiologists had a recall rate up to 57% for this group ([Table 4](#)). The radiologists had a recall rate of 2%, and the rate for abnormal assessments by AI was 3.8% in the group with normal mammograms and that with benign biopsy findings ([Table 4](#)). For the group with only normal mammograms, the recall rates were 1% and 3.6%, respectively. Radiologists and the AI program had similar rates of recall for the total study population.

The 2 screening mammograms shown in [Figures 2](#) and [3](#) have been assessed by radiologists and the AI cancer detection program. These examples illustrate concordant and discordant assessments.

Table 2. Artificial intelligence (AI) score for each age group of the normal, benign biopsy, and cancer groups.

Age group (years)	Normal mammography		Benign biopsy group						Cancer group					
	All groups (P=.40)		All groups (P=.71)		Benign biopsy findings before mammography (P=.71)		Benign biopsy findings after mammography (P=.81)		All groups (P=.01)		Prior benign biopsy findings (P=.54)		No benign biopsy findings (P=.01)	
	Participants, n	AI score, median (IQR)	Participants, n	AI score, median (IQR)	Participants, n	AI score, median (IQR)	Participants, n	AI score, median (IQR)	Participants, n	AI score, median (IQR)	Participants, n	AI score, median (IQR)	Participants, n	AI score, median (IQR)
All	9738	0.018 (0.005-0.065)	234	0.051 (0.016-0.174)	64	0.069 (0.018-0.179)	170	0.047 (0.016-0.205)	917	0.933 (0.667-0.983)	20	0.924 (0.747-0.986)	897	0.934 (0.666-0.983)
40-49	3152	0.021 (0.006-0.068)	144	0.048 (0.016-0.155)	43	0.074 (0.019-0.142)	101	0.042 (0.012-0.176)	213	0.861 (0.272-0.974)	9	0.861 (0.286-0.974)	204	0.864 (0.264-0.975)
50-59	2834	0.015 (0.005-0.058)	68	0.051 (0.017-0.260)	16	0.039 (0.014-0.113)	52	0.051 (0.017-0.305)	228	0.948 (0.715-0.985)	3	0.911 (0.891-0.950)	225	0.949 (0.709-0.985)
60-69	2577	0.018 (0.005-0.064)	22	0.078 (0.012-0.174)	5	0.173 (0.090-0.449)	17	0.047 (0.011-0.144)	379	0.935 (0.723-0.984)	8	0.945 (0.887-0.991)	371	0.935 (0.720-0.984)
≥70	1175	0.020 (0.006-0.077)	0	N/A ^a	0	N/A	0	N/A	97	0.958 (0.820-0.985)	0	N/A	97	0.958 (0.820-0.985)

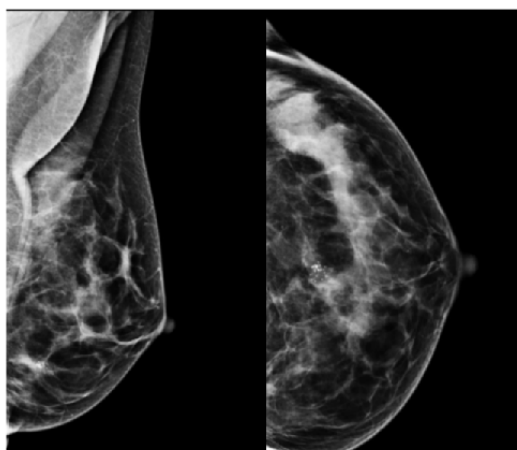
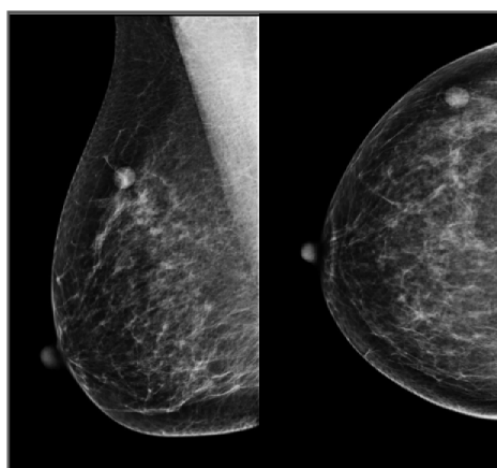
^aN/A: not applicable.

Table 3. Mammographic examinations of women a benign biopsy finding having an artificial intelligence (AI) score above the predefined threshold for cancer suspicion, grouped by the timing of the biopsy.

Timing of biopsy	Participants, n	Age (years), median (IQR)	Proportion of assessments above the cut-off point, % (n/n)	AI score	P value
Benign biopsy finding before mammography (months)	64				
0-6	9	49.1 (44.5-54.9)	33 (3/9)	0.150 (0.099-0.449)	Reference
6-24	39	48.1 (46.7-52.7)	5.1 (2/39)	0.062 (0.016-0.150)	.12
>24	16	41.0 (40.3-48.4)	0 (0)	0.025 (0.015-0.099)	.06
Benign biopsy after mammography (months)	170				
0-6	104	48.3 (44.4-52.5)	16 (17/104)	0.065 (0.017-0.274)	Reference
6-24	44	49.0 (45.6-55.3)	4.5 (2/44)	0.037 (0.008-0.109)	.36
>24	22	51.2 (45.6-53.0)	9 (2/22)	0.031 (0.017-0.165)	.38

Table 4. Recall rate and abnormal assessments by artificial intelligence.

Assessments	Recall rate, % (n/n)			Abnormal AI assessments above the cutoff point
	Radiologist 1	Radiologist 2	Consensus	
Total	10 (831/8307)	10 (827/8307)	9.2 (767/8307)	11 (880/8307)
Normal and benign biopsy findings	3.6 (274/7583)	3.1 (234/7583)	2 (154/7583)	3.8 (290/7583)
Normal	2.8 (203/7371)	2.2 (162/7371)	1 (73/7371)	3.6 (265/7371)
Benign biopsy findings	33 (71/212)	34 (72/212)	38 (81/212)	12 (25/212)
Biopsy before mammography	8.5 (5/59)	6.8 (4/59)	6.8 (4/59)	8.5 (5/59)
Biopsy after mammography	49 (66/135)	50 (68/135)	57 (77/135)	15 (20/135)
Cancer	77 (557/724)	82 (593/724)	85 (613/724)	81 (590/724)
With benign biopsy findings	75 (12/16)	88 (14/16)	88 (14/16)	81 (13/16)
Without benign biopsy findings	77 (545/708)	82 (579/708)	85 (599/708)	82 (577/708)

Figure 2. A 50-year-old woman selected by radiologists for potential pathology in the left breast. A high artificial intelligence (AI) score was assigned. The biopsy results showed hyperplastic breast epithelial cells that could represent a degenerated fibroadenoma.**Figure 3.** A 56-year-old woman selected by radiologists for potential pathology in the right breast. A low artificial intelligence (AI) score was assigned. The biopsy shows the lymph node.

Discussion

The AI CAD system in this study showed increased flagging of screening examinations for women with benign biopsy findings compared to those for healthy women without biopsies.

However, the flagging rate was similar between AI and radiologists for women with a prior biopsy finding, and considerably lower for women with a biopsy finding after screening.

For women with a previous benign biopsy finding, compared to healthy women, the AI CAD system's flagging rate (false positives) increased from 3.6% to 8.5%. In other words, there was a significant difference in AI scores between the normal group and the benign biopsy group despite both groups consisting of women without breast cancer. This finding might raise questions about the probability that AI is affected by alternations on mammography because of the biopsy. This did not seem to be the case, since we found a similar increase in recall rate for the radiologists from 2.8% to 8.5%. This is unexpected since radiologists had access to the outcomes of prior biopsies while AI did not.

For women who had a benign biopsy finding after screening, we found that 57% of them resulted from recall by the screening radiologists. Applying the AI CAD system in screening would have resulted in a much lower false positive flagging rate of only 15% for the AI program. Based on this observation, one may suggest further research on the role of AI in reducing the number of unnecessary biopsies.

The strength of this study is the large number of women with cancer and that all women were sampled from a screening cohort. Another strength of this study is the use of the specific AI algorithm, which has already been validated in large cohorts with very positive results [9]. Our data of the total recall rates and specifically those of the cancer group amplify the indications

from previous studies that AI-based cancer detectors can be reliable enough to be incorporated in a screening setting.

The main limitation is the relatively small number of benign biopsies, which makes it difficult to consider the effect of different types of benign lesions. Another limitation is the study's retrospective setting. Since the AI program did not have the opportunity to make recalls and choose women for further diagnostic biopsy, it could not influence who received a biopsy after screening, and all decisions about benign biopsies were based on radiologists' assessments. In contrast to radiologists, the AI program calculates a score for the likelihood of breast cancer based on the image alone and does not consider any information about symptoms given by the woman at screening.

Furthermore, in this study, we did not consider the exact location of the presumed abnormality where the AI program revealed a high AI score. Further analysis of the data can be valuable to evaluate whether the lesions that AI showed responded to the actual finding that the patient was recalled for.

In conclusion, the tested AI CAD system had an increased flagging rate of 8.5% for women with a prior benign biopsy finding; this rate was not higher than that for radiologists who often have information about prior benign biopsy findings. Further research and development might be focused on how to further improve AI CAD systems by taking into account information about prior benign biopsy findings.

Conflicts of Interest

FS received speaker fees from Lunit and Pfizer and is a shareholder in ClearScanAI.

Multimedia Appendix 1

Artificial intelligence (AI) score distribution for the cancer group (n=917).

[[DOCX File , 16 KB - ai_v2i1e48123_app1.docx](#)]

Multimedia Appendix 2

Artificial intelligence (AI) score distribution for the benign biopsy group (n=234).

[[DOCX File , 16 KB - ai_v2i1e48123_app2.docx](#)]

Multimedia Appendix 3

Artificial intelligence (AI) score distribution for the normal group (n=9738).

[[DOCX File , 16 KB - ai_v2i1e48123_app3.docx](#)]

References

1. Wigzell K, Rosen M. Socialstyrelsen - The National Board of Health and Welfare/Centre for Epidemiology - Foreword. *Scand J Public Health* 2001;1.
2. Ferlay J, Soerjomataram I, Dikshit R, Eser S, Mathers C, Rebelo M, et al. Cancer incidence and mortality worldwide: sources, methods and major patterns in GLOBOCAN 2012. *Int J Cancer* 2015 Mar 01;136(5):E359-E386 [[FREE Full text](#)] [doi: [10.1002/ijc.29210](https://doi.org/10.1002/ijc.29210)] [Medline: [25220842](https://pubmed.ncbi.nlm.nih.gov/25220842/)]
3. 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 04;71(3):209-249 [[FREE Full text](#)] [doi: [10.3322/caac.21660](https://doi.org/10.3322/caac.21660)] [Medline: [33538338](https://pubmed.ncbi.nlm.nih.gov/33538338/)]
4. Nyström L, Rutqvist LE, Wall S, Lindgren A, Lindqvist M, Rydén S, et al. Breast cancer screening with mammography: overview of Swedish randomised trials. *Lancet* 1993 Apr 17;341(8851):973-978. [doi: [10.1016/0140-6736\(93\)91067-v](https://doi.org/10.1016/0140-6736(93)91067-v)] [Medline: [8096941](https://pubmed.ncbi.nlm.nih.gov/8096941/)]

5. Marmot MG, Altman DG, Cameron DA, Dewar JA, Thompson SG, Wilcox M. The benefits and harms of breast cancer screening: an independent review. *Br J Cancer* 2013 Jun 11;108(11):2205-2240 [[FREE Full text](#)] [doi: [10.1038/bjc.2013.177](https://doi.org/10.1038/bjc.2013.177)] [Medline: [23744281](#)]
6. Long H, Brooks JM, Harvie M, Maxwell A, French DP. Correction: How do women experience a false-positive test result from breast screening? A systematic review and thematic synthesis of qualitative studies. *Br J Cancer* 2021 Sep 30;125(7):1031-1031 [[FREE Full text](#)] [doi: [10.1038/s41416-021-01503-w](https://doi.org/10.1038/s41416-021-01503-w)] [Medline: [34331024](#)]
7. Rodríguez-Ruiz A, Krupinski E, Mordang J, Schilling K, Heywang-Köbrunner SH, Sechopoulos I, et al. Detection of breast cancer with mammography: effect of an artificial intelligence support system. *Radiology* 2019 Feb;290(2):305-314. [doi: [10.1148/radiol.2018181371](https://doi.org/10.1148/radiol.2018181371)] [Medline: [30457482](#)]
8. Schaffter T, Buist DSM, Lee CI, Nikulin Y, Ribli D, Guan Y, the DM DREAM Consortium, et al. Evaluation of combined artificial intelligence and radiologist assessment to interpret screening mammograms. *JAMA Netw Open* 2020 Mar 02;3(3):e200265 [[FREE Full text](#)] [doi: [10.1001/jamanetworkopen.2020.0265](https://doi.org/10.1001/jamanetworkopen.2020.0265)] [Medline: [32119094](#)]
9. Salim M, Wählin E, Dembrower K, Azavedo E, Foukakis T, Liu Y, et al. External evaluation of 3 commercial artificial intelligence algorithms for independent assessment of screening mammograms. *JAMA Oncol* 2020 Oct 01;6(10):1581-1588 [[FREE Full text](#)] [doi: [10.1001/jamaoncol.2020.3321](https://doi.org/10.1001/jamaoncol.2020.3321)] [Medline: [32852536](#)]
10. Rodríguez-Ruiz A, Lång K, Gubern-Merida A, Teuwen J, Broeders M, Gennaro G, et al. Can we reduce the workload of mammographic screening by automatic identification of normal exams with artificial intelligence? A feasibility study. *Eur Radiol* 2019 Sep 16;29(9):4825-4832 [[FREE Full text](#)] [doi: [10.1007/s00330-019-06186-9](https://doi.org/10.1007/s00330-019-06186-9)] [Medline: [30993432](#)]
11. Dembrower K, Wählin E, Liu Y, Salim M, Smith K, Lindholm P, et al. Effect of artificial intelligence-based triaging of breast cancer screening mammograms on cancer detection and radiologist workload: a retrospective simulation study. *Lancet Digit Health* 2020 Sep;2(9):e468-e474. [doi: [10.1016/s2589-7500\(20\)30185-0](https://doi.org/10.1016/s2589-7500(20)30185-0)]
12. Rodríguez-Ruiz A, Lång K, Gubern-Merida A, Broeders M, Gennaro G, Clauser P, et al. Stand-alone artificial intelligence for breast cancer detection in mammography: comparison with 101 radiologists. *J Natl Cancer Inst* 2019 Sep 01;111(9):916-922 [[FREE Full text](#)] [doi: [10.1093/jnci/djy222](https://doi.org/10.1093/jnci/djy222)] [Medline: [30834436](#)]
13. Dumitrescu RG, Cotarla I. Understanding breast cancer risk -- where do we stand in 2005? *J Cell Mol Med* 2005 Jan;9(1):208-221 [[FREE Full text](#)] [doi: [10.1111/j.1582-4934.2005.tb00350.x](https://doi.org/10.1111/j.1582-4934.2005.tb00350.x)] [Medline: [15784178](#)]
14. Hartmann LC, Sellers TA, Frost MH, Lingle WL, Degnim AC, Ghosh K, et al. Benign breast disease and the risk of breast cancer. *N Engl J Med* 2005 Jul 21;353(3):229-237. [doi: [10.1056/nejmoa044383](https://doi.org/10.1056/nejmoa044383)]
15. Barlow W, White E, Ballard-Barbash R, Vacek P, Titus-Ernstoff L, Carney P, et al. Prospective breast cancer risk prediction model for women undergoing screening mammography. *J Natl Cancer Inst* 2006 Sep 06;98(17):1204-1214. [doi: [10.1093/jnci/djj331](https://doi.org/10.1093/jnci/djj331)] [Medline: [16954473](#)]
16. McCann J, Stockton D, Godward S. Impact of false-positive mammography on subsequent screening attendance and risk of cancer. *Breast Cancer Res* 2002 Oct 1;4(5):R11 [[FREE Full text](#)] [doi: [10.1186/bcr455](https://doi.org/10.1186/bcr455)] [Medline: [12223128](#)]
17. Dembrower K, Lindholm P, Strand F. A multi-million mammography image dataset and population-based screening cohort for the training and evaluation of deep neural networks-the Cohort of Screen-Aged Women (CSAW). *J Digit Imaging* 2020 Apr 13;33(2):408-413 [[FREE Full text](#)] [doi: [10.1007/s10278-019-00278-0](https://doi.org/10.1007/s10278-019-00278-0)] [Medline: [31520277](#)]
18. Lehman CD, Arao RF, Sprague BL, Lee JM, Buist DSM, Kerlikowske K, et al. National performance benchmarks for modern screening digital mammography: update from the Breast Cancer Surveillance Consortium. *Radiology* 2017 Apr;283(1):49-58 [[FREE Full text](#)] [doi: [10.1148/radiol.2016161174](https://doi.org/10.1148/radiol.2016161174)] [Medline: [27918707](#)]
19. Yala A, Schuster T, Miles R, Barzilay R, Lehman C. A deep learning model to triage screening mammograms: a simulation study. *Radiology* 2019 Oct;293(1):38-46. [doi: [10.1148/radiol.2019182908](https://doi.org/10.1148/radiol.2019182908)] [Medline: [31385754](#)]

Abbreviations

- AI:** artificial intelligence
 - CAD:** computer-aided detection
 - CSAW:** Cohort of Screen-Aged Women
 - RCC:** Regional Cancer Center
-

Edited by K El Emam, G Eysenbach; submitted 12.04.23; peer-reviewed by A Karakatsanis, N Mungoli; comments to author 07.05.23; revised version received 17.06.23; accepted 03.08.23; published 31.08.23.

Please cite as:

Zouzos A, Milovanovic A, Dembrower K, Strand F

Effect of Benign Biopsy Findings on an Artificial Intelligence–Based Cancer Detector in Screening Mammography: Retrospective Case-Control Study

JMIR AI 2023;2:e48123

URL: <https://ai.jmir.org/2023/1/e48123>

doi: [10.2196/48123](https://doi.org/10.2196/48123)

PMID: [38875554](https://pubmed.ncbi.nlm.nih.gov/38875554/)

©Athanasios Zouzos, Aleksandra Milovanovic, Karin Dembrower, Fredrik Strand. Originally published in JMIR AI (<https://ai.jmir.org>), 31.08.2023. This is an open-access article distributed under the terms of the Creative Commons Attribution License (<https://creativecommons.org/licenses/by/4.0/>), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR AI, is properly cited. The complete bibliographic information, a link to the original publication on <https://www.ai.jmir.org/>, as well as this copyright and license information must be included.

Viewpoint

Artificial Intelligence Algorithms in Health Care: Is the Current Food and Drug Administration Regulation Sufficient?

Meghavi Mashar^{1*}, MBBCHIR, MA; Shreya Chawla^{2*}, MBBS; Fangyue Chen³, MBBCHIR, MA; Baker Lubwama⁴, BA; Kyle Patel⁵, BA; Mihir A Kelshiker⁶, MBBS; Patrik Bachtiger⁶, MBBS; Nicholas S Peters⁶, MD

¹University College London NHS Foundation Trust, London, United Kingdom

²Faculty of Life Sciences and Medicine, King's College of London, London, United Kingdom

³School of Public Health, Faculty of Medicine, Imperial College London, London, United Kingdom

⁴School of Clinical Medicine, University of Cambridge, Cambridge, United Kingdom

⁵Sloan School of Management, Massachusetts Institute of Technology, Cambridge, MA, United States

⁶National Heart and Lung Institute, Faculty of Medicine, Imperial College London, London, United Kingdom

*these authors contributed equally

Corresponding Author:

Fangyue Chen, MBBCHIR, MA

School of Public Health

Faculty of Medicine

Imperial College London

Level 2, Faculty Building South Kensington Campus

London, SW72AZ

United Kingdom

Phone: 44 7510888677

Email: fangyue.chen@nhs.net

Abstract

Given the growing use of machine learning (ML) technologies in health care, regulatory bodies face unique challenges in governing their clinical use. Under the regulatory framework of the Food and Drug Administration, approved ML algorithms are practically locked, preventing their adaptation in the ever-changing clinical environment, defeating the unique adaptive trait of ML technology in learning from real-world feedback. At the same time, regulations must enforce a strict level of patient safety to mitigate risk at a systemic level. Given that ML algorithms often support, or at times replace, the role of medical professionals, we have proposed a novel regulatory pathway analogous to the regulation of medical professionals, encompassing the life cycle of an algorithm from inception, development to clinical implementation, and continual clinical adaptation. We then discuss in-depth technical and nontechnical challenges to its implementation and offer potential solutions to unleash the full potential of ML technology in health care while ensuring quality, equity, and safety. References for this article were identified through searches of PubMed with the search terms “Artificial intelligence,” “Machine learning,” and “regulation” from June 25, 2017, until June 25, 2022. Articles were also identified through searches of the reference list of the articles. Only papers published in English were reviewed. The final reference list was generated based on originality and relevance to the broad scope of this paper.

(JMIR AI 2023;2:e42940) doi:[10.2196/42940](https://doi.org/10.2196/42940)

KEYWORDS

artificial intelligence; machine learning; regulation

Introduction

Machine learning (ML) technology aims to improve the quality and efficiency of health care within the current health systems. Its applications encompass roles traditionally undertaken by health care professionals, such as clinical triage at emergency departments, mammography screening, and diagnosis undertaken by radiologists [1,2]. In many studies, ML algorithms

have outperformed clinicians, for instance, in chest radiograph interpretation, skin cancer diagnosis, and directing optimal treatment strategies for sepsis in intensive care [3,4].

ML-based adaptive algorithms have the ability to *learn* and optimize their performance within the ever-changing clinical environment. The adaptability helps to optimize its clinical

utility but has the potential to impact patient safety by introducing an element of unpredictability.

While there has been a significant increase in the volume of literature describing ML since 2010 [5], the regulation of adaptive ML technology has lagged behind its rapid technological advancement. In the United States, the current framework under the Food and Drug Administration (FDA) only regulates an algorithm at the point of clinical deployment but fails to account for the initial model inception, development, and evolution once deployed into clinical use. In the United Kingdom, the National Health Service (NHS) has accelerated its effort in digitalization within health care through the creation of NHSx and NHS AI Lab, with an emphasis on the development of a suitable governance framework for artificial intelligence (AI) in health care [6]. Elsewhere in the world, ML regulation is at varying stages. India does not draw a distinction between ML algorithms and other medical devices, while China's New Generation Artificial Intelligence Development Plan does not address regulation of medical devices [7,8]. While the World Health Organization has published guiding principles for ML use, it does not outline a specific framework for regulation [9].

This paper aims to use the current FDA regulatory model as an example, build on the existing framework, and propose a novel regulatory pathway for ML algorithms from inception through clinical deployment to model evolution. Since ML algorithms aim to support or, in certain cases, replace the role of medical professionals, we likened the regulatory pathway to those of medical professionals. We then discuss the associated challenges to its implementation and potential solutions to overcome the challenges.

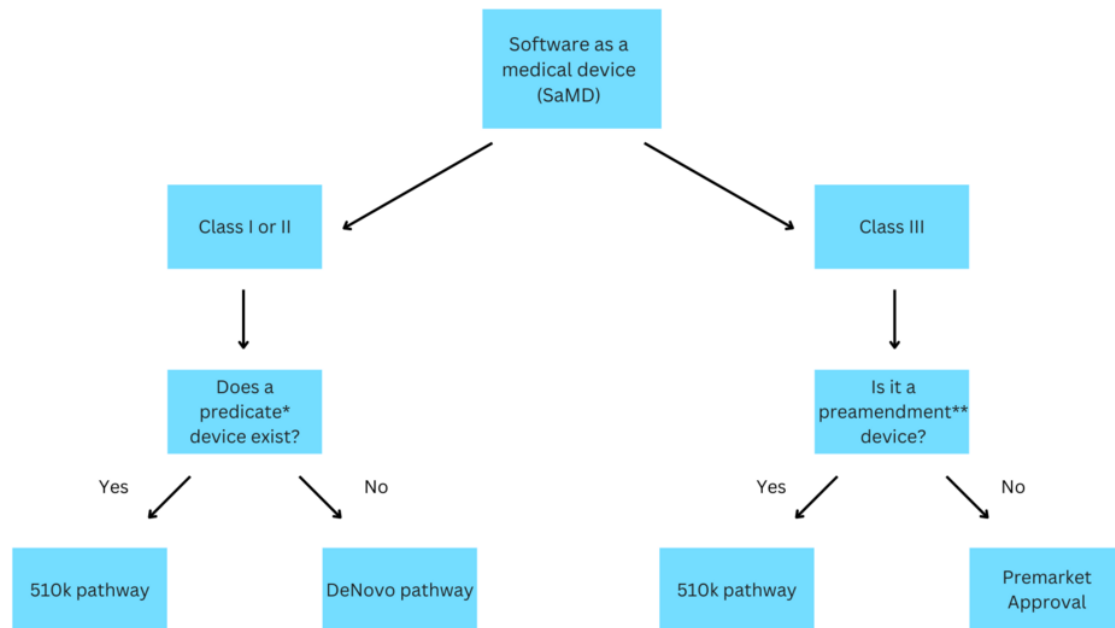
Current Regulatory Pathways and Potential Issues

Currently, most ML algorithms are approved by the FDA through one of three pathways: 510k, premarket approval, or the DeNovo pathway (see Figure 1) [10-12]. At a single timepoint prior to its approval, the ML production company will need to demonstrate the safety and effectiveness of the algorithm within its intended use. The current benchmark for approval requires companies to demonstrate good model performance on a varied data set and in a real-world setting. With no explicit definition of what constitutes reproducible standards, it is no surprise that the current FDA-approved ML algorithms vary considerably by the size of data sets and number of sites [5].

Under the current regulation, once an algorithm is approved, its behavior will remain fixed, defeating the distinguishing advantage of many adaptive algorithms in their ability to *learn* throughout their life cycle. Its current inflexible state not only reduces its clinical utility but can alarmingly infringe patient safety. For instance, an algorithm trained in 2018 to recognize pneumonia on a chest radiograph will not be able to differentiate it from COVID-19. Furthermore, variations exist in disease prevalence and population demographics across sites, such that the internal training and testing data sets used during algorithm development may not be representative of the population they are deployed to, thus performing poorly during external validation [13-15]. Moreover, depending on the training data set, the model may not be able to respond to geographically, ethnically, and socioeconomically diverse patient cohorts.

Additionally, the current FDA framework does not regulate the inception of an ML algorithm. As a result, a number of algorithms have been approved, many of similar use cases with varying development sites and data sizes. This can potentially constitute an inefficient use of resources [16].

Figure 1. The current Food and Drug Administration (FDA) regulatory pathway. *A predicate: if the algorithm is found to be substantially equivalent to a legally marketed device. **A preamendment device: devices legally marketed in the United States before May 28, 1976, which have not been significantly changed/modified and for which no regulation requiring premarket approval has been published by the FDA.



Current Attempts to Support Model Evolution

In April 2018, to account for the iterative improvement in ML model performance as new training data and improved data science techniques become available, the FDA released a white paper outlining a proposed framework for the regulation of ML-based software in medicine [17].

The proposed Total Product Lifecycle (TPLC) regulatory approach allows for iterative product improvement while maintaining essential safeguards. The framework adopts the principle of a *Predetermined Change Control Plan* produced by the manufacturer, which aims to anticipate potential modifications during clinical deployment. The Software as a Medical Device Pre-Specifications (SPS) will underline the modification expected by the manufacturer relating to *performance, inputs, and intended use*. Modifications within the SPS can be implemented without the need to resubmit for marketing application.

The implementation of the TPLC approach thus places the onus on the manufacturers to monitor and evaluate algorithm performance during its clinical use and regularly report to the FDA with updates and performance metrics. The culture of quality and organizational excellence of the company would be assessed according to the outlined standards in Good Machine-Learning Practice (GMLP). To date, only a single manufacturer of a cardiac ultrasound software has used the *Predetermined Change Control Plan* to facilitate future model alterations [18].

Elsewhere, similar trends have been observed in ML regulatory policies. The European Union has recently introduced the EU Medical Device Regulation, imposing stringent regulatory requirements from early-stage considerations through algorithm

development to postmarket surveillance that need to be met prior to the clinical use of medical devices, including ML algorithms [19]. Likewise, in the United Kingdom, a code of conduct for AI and data-driven technology has been introduced to facilitate collaboration between technological companies and the NHS in developing high-quality safe medical devices [20].

While the TPLC approach has set out a useful theoretical framework in addressing the adaptive nature of ML algorithms, it places heavy emphasis on the manufacturer in governing the algorithm post deployment and overlooks the need to involve local end users immersed in the clinical environment. Moreover, despite being proposed for some time, the TPLC framework is yet to be implemented, which likely stems from the complexities involved. The framework also does not accommodate the evolution of algorithms beyond the predetermined specifications and change protocols. Finally, the framework has not addressed wider issues of the clinical utility, data suitability, and health equity of ML algorithms, which may call for a greater degree of regulation at a much earlier stage in the model life cycle.

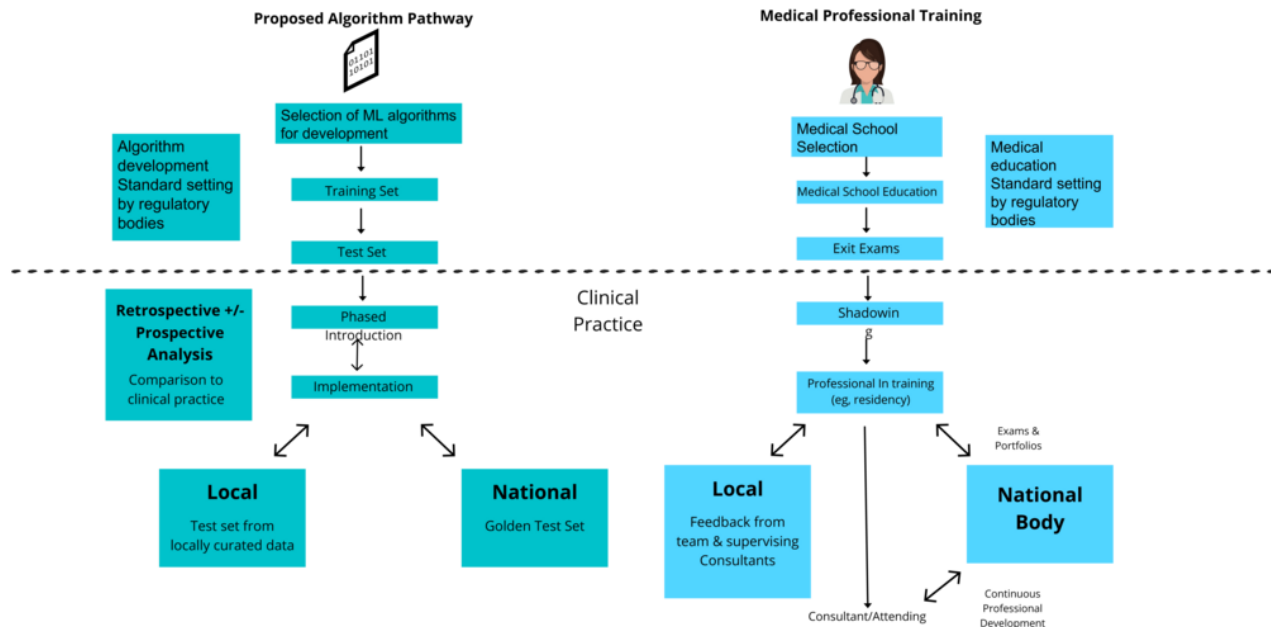
In January 2021, the FDA released a document outlining an action plan in response to feedback from stakeholders on the TPLC approach, as well as SPS and the *Predetermined Change Control Plan* [21]. The five-point plan expressed the FDA's intention to facilitate various enhancements to their TPLC approach, such as furthering GMLP by participating in communities (eg, the Xavier AI World Consortium) that collaborate to promote best practices in ML. In addition, the document expressed an appreciation for the need for a *patient-centered approach* as well as the evaluation of real-world model performance. The current action taken includes working with volunteers and engaging in further research to consider methods for real-world performance monitoring. Therefore, while the FDA has acknowledged many of the stakeholder queries (including some mentioned in this paper), there still is not much in the way of tangible solutions.

The Proposed Regulatory Pathway

Currently, the regulation of ML algorithms is akin to those for drug development [22]. However, the lack of ongoing prospective evaluation of AI algorithms truly limits their use

in practice. As such, ML algorithms share a greater analogy to medical professionals, as they often undertake or support tasks traditionally performed by them and are subject to ongoing regulation. We therefore propose an analogous regulatory framework for ML algorithms, as summarized in Figure 2.

Figure 2. The proposed algorithm regulatory pathway analogous to the current medical professional training pathway. ML: machine learning.



At the start, aspiring medical professionals are required to go through a selective process that ensures their baseline capabilities and suitability to begin their medical education. Similarly, the inception of an ML algorithm begins with a clinical problem that it aims to solve in health care. Algorithms across health care fields should be contested on their clinical value, usability, cost-effectiveness, and sustainability prior to its development, which will help to direct resources appropriately.

The model development phase can be likened to the undergraduate training of medical professionals. In the United Kingdom, the General Medical Council sets out standards and expected outcomes for medical education across the 44 recognized medical schools [23]. Similarly, in the face of the current heterogeneity present in the approved ML algorithms, structured standard-setting by an independent regulatory body should be in place during the development of algorithms on indicators such as data size and quality, technical assurance, and clinical safety. Guidelines such as TRIPOD-AI (Transparent Reporting of a Multivariable Prediction Model for Individual Prognosis or Diagnosis–Artificial Intelligence) and PROBAST-AI (Prediction Model Risk of Bias Assessment Tool–Artificial Intelligence) are being developed to help appraise AI-based prediction and diagnostic models [24]. These can be incorporated into the regulatory framework.

Prior to clinical deployment, the final clinical efficacy of ML algorithms is determined by a test data set, akin to the exit examination undertaken by medical professionals prior to qualification and employment. We propose further stages after the current regulation that ends after clinical deployment of ML algorithms.

Medical professionals often enter a period of supervision prior to independent practice, for instance, the internship period (foundation program) in the United Kingdom, which allows them to adapt to clinical practice [25]. Similarly, we propose that ML algorithms should enter a period of phased introduction that will involve an initial trial period for the algorithm to *observe*, operate alongside clinicians, and adjust to local working practices and systems. Ongoing evaluation and adaptation will take place in preparation for its full deployment.

After the initial period of shadowing, medical professionals are continuously re-evaluated to demonstrate ongoing competencies at a local and national level through national body board examinations [26], continuing professional development, and clinical portfolios [27,28]. We propose analogous local and national regulations for ML algorithms to ensure that they are consistently pertinent and useful in the ever-changing landscape of clinical practice. Locally, we propose for institutions to curate their own test sets containing representative cases that better reflect the variability in equipment, protocol, epidemiology, and patient populations encountered at the deployment site. Should the local testing demonstrate deficiencies, models can then be retrained on the curated local training data sets.

During the local training process, it is imperative that the algorithm does not deviate substantially from its initial objectives and continues to provide its proposed clinical benefit. While the FDA has delegated the task of ongoing data collection and monitoring of the algorithms to the manufacturer, a dedicated national regulatory body may be more suitable for this role. We thus propose the formation of national governance structures consisting of a panel of appointed experts who would be responsible for the selection of a series of cases that would

typify the minimum standard the algorithm is expected to achieve within its specified case use—known as the *golden* test set. Unlike the local regulatory bodies, the core aim is to maintain safety and basic competence of the algorithm rather than its optimization. Additionally, the national *golden* test set will be updated in response to large changes in clinical practice by selecting cases from local representative data sets, for instance, when more effective treatment emerges, such as the use of mechanical thrombectomy for patients with stroke [29]; changes in policies, such as radiology imaging guidelines; and changes in pathology, such as the COVID-19 emergency.

The Complexity Behind ML Regulation: Now and the Future

Both technical and nontechnical barriers pose a challenge to the implementation of any effective regulatory model. This may also explain why the TPLC approach has not been implemented more than 3 years following its inception. Ongoing regulation of an ML algorithm requires mechanisms to monitor model performance and methods of updating the model, the latter necessitating data sharing.

Facilitating Model Evolution

Model drift, a process where the model's prediction power deteriorates due to changes in the clinical environment, is the main cause of deviation in model performance once deployed clinically. The proposed regulatory pathway aims to engineer a performance monitoring and adaptation system on a local and national level that aims to detect, monitor, and mitigate the effect of model drift. Logistically, this process can be more nuanced.

In some circumstances, model drift can be anticipated, enabling retraining in advance of its occurrence. This is typically limited to foreseeable changes that alter the data distribution, such as a newly acquired computed tomography (CT) scanner that enables thinner reconstruction of images (eg, 1-mm thickness slices rather than 5-mm thickness). When the data in the domain is expected to change frequently, the identification of model drift can be automated so that the model can be retrained accordingly on a regular basis, both of which will require overhead infrastructure to be in place [30].

In other cases, once model drift is detected, its cause must be understood to take appropriate action. These range from biological factors such as a change in the characteristics of the patient population or management guidelines, technological factors such as novel treatment and imaging technology, and operational factors such as a change in the format of incoming data (eg, when the oxygen saturation probe outputs saturation as an integer ["97"] rather than a string ["97%"]).

To retrain ML algorithms, a wide range of methods are available from simple calibration to full retraining with the possible addition of new features. The choice of using old or new data for retraining depends on the application of the algorithm. For instance, if a specific cause has resulted in model drift, such as the above example of a novel CT scanner, then the model will need to be retrained on the new data as they are generated. If

the drift is infrequent, both data sets can be combined to update the existing algorithm or generate a new algorithm. If the data is highly dynamic, retraining can be performed on new data while replacing the old.

During the retraining process, one must strive for a fine balance between maintaining the algorithm's original function while adapting to its new local environment and minimizing new bias. For instance, in approaches that disregard the old data or algorithm, a risk of overfitting is present such that the algorithm may lose its original function. At the same time, care must be taken not to bias the model toward the outliers in the data set. For instance, the addition of a new data set with more cases of malignant chest nodules may bias an algorithm to predict lung cancer rather than benign modules from chest radiographs.

Ground-Truthing

During the local and national testing and retraining, ground-truthing, or annotation of data to compare with algorithm predictions, is an essential process during model evolution. While fully automated methods exist, typically achievable in binary classification tasks with well-structured data, more complex tasks such as segmentation tasks (eg, identification of a lesion on a scan) will require manual labeling by a specialist. The question remains as to who, when, and how this process will take place alongside the clinical workflow. Independent companies that specialize in data annotation and ground-truthing exist, which may help to circumvent this added layer of complexity.

Data Sharing

Insufficient sample size or restricted data sets can make it difficult for data to be interpreted through ML techniques subsequently introducing bias and underestimation of minority groups [31]. For example, the International Skin Imaging Collaboration: Melanoma Project, one of the largest dermatology data sets of pigmented lesions, largely focuses on Caucasian populations, which will limit its performance in other populations. Moreover, health outcomes are known to be worse in minority populations. Thus, it continues to be imperative to be able to acquire a range of data from a variety of sources to train ML models [32].

However, data sharing poses a challenge due to the sensitive nature of patient data and the sheer volume of data to be transferred [33]. During the current workflow for the development of ML algorithms, clinical sites typically share medical data for a specified period of time through two pathways: direct sharing and data enclaves. The former involves sending data out of the clinical network to the developers, while the latter takes the opposite approach by allowing external model developers into the clinical sites. Both routes can open up the potential for data misuse outside the agreed terms and compromise patient trust and safety. Data curated across multiple sites help to improve algorithm performance and minimize bias but will require greater stringency in its governance and standardization.

One potential solution is federated learning, a process that allows a model to be trained on multiple data sets across different sites by solely allowing access to specific features of each data set

without physically exchanging data [34]. This circumvents the risks of data sharing while increasing the size and diversity of case pathology and demographics the algorithm is exposed to. Moreover, federated learning opens up the possibility of continuous learning by ongoing access to live data, rather than the outdated static data sets procured through the current two pathways. Collectively through a federated platform, the performance of the algorithm can be constantly tracked, trained, and tested.

Nevertheless, to unlock the full potential of this technique, we will need to overcome several logistical challenges. First, the initial algorithm development will still require intimate access to data. Second, data across sites can be stored in variable formats, making it more difficult to standardize and access the specified features required for federated learning. Finally, federated learning will need to be supported by adequate local hardware and networks, and can be bottlenecked by resource-constrained sites [34].

A number of ML- and non-ML-based prediction tools have been developed using national and international collaborative data sets [35-37]. In Taiwan, the National Health Insurance Research Database exemplifies a population-level data source for research in health care, with strict requirement for privacy and data confidentiality [38]. The Chronic Kidney Disease Prognosis Consortium, international collaborative data sets sponsored by the US National Kidney Foundation, harnesses data from over 80 population cohorts in an effort to improve the global outcome of kidney disease [39]. The use of data often requires stringent application through research institutions and public bodies. This, however, helps to optimize data quality, size, and diversity in a collaborative effort to direct ML technology toward priority areas while ensuring an optimal level of data governance.

Integration Into Clinical Practice

Ultimately, the approved algorithms will need to yield sufficient clinical value to be accepted and integrated into the existing clinical workflow. Medical professionals will need to adapt their clinical practice and maximize the utility of the new technology. At the same time, ML algorithms make mistakes, as exemplified by the erroneous treatment recommendations made by IBM Watson for Oncology and the more recent Epic Sepsis Model that was found to miss two-thirds of sepsis cases that it was designed to predict [40,41]. Astringent safeguarding processes should be put in place, as the risk of faulty algorithms can affect a population at a system level, rather than of a single doctor-patient interaction [3].

Adaptation of Medical Professionals

The introduction of ML algorithms into the clinical workflow of medical professionals will not be an easy task. As mentioned above, we propose for a period of shadow deployment of the ML algorithm to allow clinicians to acclimatize to the new practice and troubleshoot for any issues while ensuring the algorithm is safe and reliable. During its clinical practice, once an algorithm is retrained, its functions and iterations may differ, while clinicians may continue to practice based on the algorithm's prior behavior, introducing an element of automation

bias. Therefore, clinicians will be required to continually adapt their clinical practice alongside the ML algorithm to maintain a good standard of care. Nevertheless, ongoing learning is already an integral part of medical professionals' career paths. Clinicians have in the past adapted well to system changes such as the introduction of electronic health systems, the emergence of new diseases (COVID-19 being a stark example), alongside the flexibility in working with different members of the multidisciplinary team.

Looking beyond the future, the traditional health care training curriculum will need to adapt to the evolving medical technology through the introduction of ML into the medical curriculum. In fact, universities worldwide have recognized the demand for interdisciplinary medical professionals by introducing combined medical and engineering programs [42-44]. As proposed by Panch et al [45], ML may emerge as a new medical specialty to oversee the development and clinical implementation of ML algorithms into health care.

Adaptation of the Current Workflow

Ongoing local monitoring is a necessity. This will require design of a protocol and the use of specific resources. For instance, a threshold will need to be predetermined to trigger the re-evaluation of algorithm performance at a fixed interval or when a deterioration in performance is detected. When an algorithm is suspended for retraining and evaluation, a sustainable substitute will need to be in place to maintain the standards of care prior to its reintroduction.

The development of local test sets will become an additional process alongside the usual clinical practice. As to who will undergo the process of ground-truthing, the practice of internal clinicians that regularly work with the model may be influenced by the model itself, thus introducing bias to subsequent inputs. For instance, radiologists who rely on ML algorithms to detect nodules may be less adept at their detection during the ground-truthing process. On the other hand, external clinicians may be less accustomed to the local equipment and practices. The optimal solution may involve the recruitment of a representative number of internal and external clinicians to expose the algorithm to a variation in clinical practice and minimize bias. Nevertheless, the entire process of model evolution will require a learning curve for all health care workers involved.

Adaptation of the Governing Structure

At present, the FDA places the onus on the third-party manufacturers to develop, monitor, and evaluate their ML algorithms. This is no longer sufficient or efficient. As described above, independent local and national governing structures involving multiple stakeholders will need to be in place, taking on a strong oversight in regulating the development of algorithms, clinical implementation, detection of deviation in algorithm performance, curation of local and national data sets, and circumventing automation bias, all within the constraints of limited clinical resources. The governing responsibility should be shared among clinicians, managers, software engineers, parent company representatives, and patients.

Adaptation of the Health System

All local sites are not created equal. Smaller resource-limited hospitals with limited infrastructure or expertise may in fact benefit the most if the full potential of ML technology is used appropriately, supporting limited workforce resources, inefficient workflow, and inadequate time between patients and clinicians. These hospitals, however, will require extensive support. In addition, the potential increase in workload to facilitate the local evolution and monitoring of algorithms may be particularly taxing for smaller peripheral hospitals, potentially nullifying the local uptake of ML technology. Potential solutions may be in the form of a network of external ML experts as well as specialist hardware and software to support local implementation of ML algorithms, their monitoring, and evaluation. In addition, regulatory frameworks worldwide should emphasize the importance of equity and accessibility in the development of ML algorithms, taking into consideration resource-limited hospitals and countries, optimizing the use of available resources while optimizing the performance of the ML algorithms.

Conclusion

The growing use and development of ML algorithms worldwide mandate the need for robust regulatory mechanisms. Current pathways proposed by the FDA demonstrate limited scope for the algorithm to adapt to the ever-changing clinical landscape. While propositions have been made on how to improve the existing pathways, they do not involve major stakeholders and face many challenges to implementation. Given the supporting role of ML algorithms alongside medical professionals, this paper has proposed a parallel regulatory pathway from inception to implementation that allows continuous model evolution throughout its clinical course. Complexities and barriers do exist in its implementation. Successful implementation will necessitate novel, robust, and ML-specific infrastructure and governing bodies. Concomitantly, adaptability of medical professionals and interdisciplinary collaboration will be vital to unleash the full potential of ML technology in health care while ensuring quality, equity, and safety.

Authors' Contributions

MM, SC, FC, BL, and KP contributed to the conceptualization, writing of the original draft, and the review and editing of the manuscript. SC and FC prepared and finalized all figures. MAK and PB contributed to the review and editing of the manuscript. NSP supervised and oversaw the completion of the manuscript. All authors have read and approved the final manuscript.

Conflicts of Interest

None declared.

References

1. Fernandes M, Vieira SM, Leite F, Palos C, Finkelstein S, Sousa JMC. Clinical decision support systems for triage in the emergency department using intelligent systems: a review. *Artif Intell Med* 2020 Jan;102:101762. [doi: [10.1016/j.artmed.2019.101762](https://doi.org/10.1016/j.artmed.2019.101762)] [Medline: [31980099](https://pubmed.ncbi.nlm.nih.gov/31980099/)]
2. Rodríguez-Ruiz A, Krupinski E, Mordang J, Schilling K, Heywang-Köbrunner SH, Sechopoulos I, et al. Detection of breast cancer with mammography: effect of an artificial intelligence support system. *Radiology* 2019 Mar;290(2):305-314. [doi: [10.1148/radiol.2018181371](https://doi.org/10.1148/radiol.2018181371)] [Medline: [30457482](https://pubmed.ncbi.nlm.nih.gov/30457482/)]
3. Topol EJ. High-performance medicine: the convergence of human and artificial intelligence. *Nat Med* 2019 Jan;25(1):44-56. [doi: [10.1038/s41591-018-0300-7](https://doi.org/10.1038/s41591-018-0300-7)] [Medline: [30617339](https://pubmed.ncbi.nlm.nih.gov/30617339/)]
4. Komorowski M, Celi LA, Badawi O, Gordon AC, Faisal AA. The Artificial Intelligence Clinician learns optimal treatment strategies for sepsis in intensive care. *Nat Med* 2018 Nov;24(11):1716-1720. [doi: [10.1038/s41591-018-0213-5](https://doi.org/10.1038/s41591-018-0213-5)] [Medline: [30349085](https://pubmed.ncbi.nlm.nih.gov/30349085/)]
5. Benjamins S, Dhunoo P, Meskó B. The state of artificial intelligence-based FDA-approved medical devices and algorithms: an online database. *NPJ Digit Med* 2020;3:118. [doi: [10.1038/s41746-020-00324-0](https://doi.org/10.1038/s41746-020-00324-0)] [Medline: [32984550](https://pubmed.ncbi.nlm.nih.gov/32984550/)]
6. Artificial intelligence: how to get it right. NHS England: Transformation Directorate. 2019 Oct. URL: <https://www.nhs.uk/ai-lab/explore-all-resources/understand-ai/artificial-intelligence-how-get-it-right/> [accessed 2022-06-24]
7. Classification of medical device pertaining to Software under the provisions of Medical Devices Rules, 2017 - Reg. Central Drugs Standard Control Organization. URL: https://cdsco.gov.in/opencms/opencms/system/modules/CDSCO.WEB/elements/download_file_division.jsp?num_id=NzY1OQ== [accessed 2022-12-10]
8. Roberts H, Cows J, Morley J, Taddeo M, Wang V, Floridi L. The Chinese approach to artificial intelligence: an analysis of policy, ethics, and regulation. *AI Soc* 2020 Jun 17;36(1):59-77. [doi: [10.1007/s00146-020-00992-2](https://doi.org/10.1007/s00146-020-00992-2)]
9. Ethics and governance of artificial intelligence for health. World Health Organization. URL: <https://www.who.int/publications/i/item/9789240029200> [accessed 2022-12-10]
10. Premarket notification 510(k). US Food and Drug Administration. URL: <https://www.fda.gov/medical-devices/premarket-submissions-selecting-and-preparing-correct-submission/premarket-notification-510k> [accessed 2022-06-24]
11. Premarket approval (PMA). US Food and Drug Administration. URL: <https://www.fda.gov/medical-devices/premarket-submissions-selecting-and-preparing-correct-submission/premarket-approval-pma> [accessed 2022-12-10]

12. De novo classification request. US Food and Drug Administration. URL: <https://www.fda.gov/medical-devices/premarket-submissions-selecting-and-preparing-correct-submission/de-novo-classification-request> [accessed 2022-06-24]
13. Kelly CJ, Karthikesalingam A, Suleyman M, Corrado G, King D. Key challenges for delivering clinical impact with artificial intelligence. *BMC Med* 2019 Oct 29;17(1):195 [FREE Full text] [doi: [10.1186/s12916-019-1426-2](https://doi.org/10.1186/s12916-019-1426-2)] [Medline: [31665002](https://pubmed.ncbi.nlm.nih.gov/31665002/)]
14. Dexter G, Grannis S, Dixon B, Kasthurirathne S. Generalization of machine learning approaches to identify notifiable conditions from a statewide health information exchange. *AMIA Jt Summits Transl Sci Proc* 2020;2020:152-161 [FREE Full text] [Medline: [32477634](https://pubmed.ncbi.nlm.nih.gov/32477634/)]
15. Zech JR, Badgeley MA, Liu M, Costa AB, Titano JJ, Oermann EK. Variable generalization performance of a deep learning model to detect pneumonia in chest radiographs: a cross-sectional study. *PLoS Med* 2018 Nov;15(11):e1002683 [FREE Full text] [doi: [10.1371/journal.pmed.1002683](https://doi.org/10.1371/journal.pmed.1002683)] [Medline: [30399157](https://pubmed.ncbi.nlm.nih.gov/30399157/)]
16. Artificial intelligence and machine learning (AI/ML)-enabled medical devices. US Food and Drug Administration. URL: <https://www.fda.gov/medical-devices/software-medical-device-samd/artificial-intelligence-and-machine-learning-ai-ml-enabled-medical-devices> [accessed 2022-06-24]
17. Proposed regulatory framework for modifications to artificial intelligence/machine learning (AI/ML)-based software as a medical device (SaMD): discussion paper and request for feedback. US Food and Drug Administration. 2019. URL: <https://www.fda.gov/files/medical%20devices/published/US-FDA-Artificial-Intelligence-and-Machine-Learning-Discussion-Paper.pdf> [accessed 2022-12-29]
18. FDA authorizes marketing of first cardiac ultrasound software that uses artificial intelligence to guide user. US Food and Drug Administration. URL: <https://www.fda.gov/news-events/press-announcements/fda-authorizes-marketing-first-cardiac-ultrasound-software-uses-artificial-intelligence-guide-user> [accessed 2022-06-24]
19. Beckers R, Kwade Z, Zanca F. The EU medical device regulation: implications for artificial intelligence-based medical device software in medical physics. *Phys Med* 2021 Mar;83:1-8. [doi: [10.1016/j.ejmp.2021.02.011](https://doi.org/10.1016/j.ejmp.2021.02.011)] [Medline: [33657513](https://pubmed.ncbi.nlm.nih.gov/33657513/)]
20. Department of Health and Social Care. New code of conduct for artificial intelligence (AI) systems used by the NHS. GOV.UK. URL: <https://www.gov.uk/government/news/new-code-of-conduct-for-artificial-intelligence-ai-systems-used-by-the-nhs> [accessed 2022-06-24]
21. Artificial intelligence/machine learning (AI/ML)-based software as a medical device (SaMD) action plan. US Food and Drug Administration. URL: <https://www.fda.gov/media/145022/download> [accessed 2022-12-10]
22. Smallman M. Policies designed for drugs won't work for AI. *Nature* 2019 Mar;567(7746):7. [doi: [10.1038/d41586-019-00737-2](https://doi.org/10.1038/d41586-019-00737-2)] [Medline: [30842640](https://pubmed.ncbi.nlm.nih.gov/30842640/)]
23. Outcomes for graduates. General Medical Council. URL: <https://www.gmc-uk.org/education/standards-guidance-and-curricula/standards-and-outcomes/outcomes-for-graduates/outcomes-for-graduates> [accessed 2022-06-24]
24. Collins GS, Dhiman P, Andaur Navarro CL, Ma J, Hooft L, Reitsma JB, et al. Protocol for development of a reporting guideline (TRIPOD-AI) and risk of bias tool (PROBAST-AI) for diagnostic and prognostic prediction model studies based on artificial intelligence. *BMJ Open* 2021 Jul 09;11(7):e048008 [FREE Full text] [doi: [10.1136/bmjopen-2020-048008](https://doi.org/10.1136/bmjopen-2020-048008)] [Medline: [34244270](https://pubmed.ncbi.nlm.nih.gov/34244270/)]
25. Shadowing for junior doctors. British Medical Association. URL: <https://www.bma.org.uk/advice-and-support/career-progression/training/shadowing-for-junior-doctors> [accessed 2022-06-24]
26. National professional examinations. General Medical Council. URL: <https://www.gmc-uk.org/education/standards-guidance-and-curricula/guidance/national-professional-examinations> [accessed 2022-06-24]
27. Continuing professional development. General Medical Council. URL: <https://www.gmc-uk.org/education/standards-guidance-and-curricula/guidance/continuing-professional-development> [accessed 2022-06-24]
28. Performance assessments. General Medical Council. URL: <https://www.gmc-uk.org/concerns/information-for-doctors-under-investigation/performance-assessments> [accessed 2022-06-24]
29. Rodrigues FB, Neves JB, Caldeira D, Ferro JM, Costa J. Endovascular treatment versus medical care alone for ischaemic stroke: systematic review and meta-analysis. *BMJ* 2019 Apr 11;365:l1658. [doi: [10.1136/bmj.l1658](https://doi.org/10.1136/bmj.l1658)] [Medline: [30975661](https://pubmed.ncbi.nlm.nih.gov/30975661/)]
30. Žliobaitė I, Pechenizkiy M, Gama J. An overview of concept drift applications. In: Japkowicz N, Stefanowski J, editors. *Big Data Analysis: New Algorithms for a New Society*. Cham: Springer; 2015:91-114.
31. Gianfrancesco MA, Tamang S, Yazdany J, Schmajuk G. Potential biases in machine learning algorithms using electronic health record data. *JAMA Intern Med* 2018 Nov 01;178(11):1544-1547 [FREE Full text] [doi: [10.1001/jamainternmed.2018.3763](https://doi.org/10.1001/jamainternmed.2018.3763)] [Medline: [30128552](https://pubmed.ncbi.nlm.nih.gov/30128552/)]
32. Aizer AA, Wilhite TJ, Chen M, Graham PL, Choueiri TK, Hoffman KE, et al. Lack of reduction in racial disparities in cancer-specific mortality over a 20-year period. *Cancer* 2014 May 15;120(10):1532-1539. [doi: [10.1002/cncr.28617](https://doi.org/10.1002/cncr.28617)] [Medline: [24863392](https://pubmed.ncbi.nlm.nih.gov/24863392/)]
33. Rieke N, Hancox J, Li W, Milletari F, Roth HR, Albarqouni S, et al. The future of digital health with federated learning. *NPJ Digit Med* 2020;3:119. [doi: [10.1038/s41746-020-00323-1](https://doi.org/10.1038/s41746-020-00323-1)] [Medline: [33015372](https://pubmed.ncbi.nlm.nih.gov/33015372/)]
34. Li T, Sahu AK, Talwalkar A, Smith V. Federated learning: challenges, methods, and future directions. *IEEE Signal Process Mag* 2020 May;37(3):50-60. [doi: [10.1109/msp.2020.2975749](https://doi.org/10.1109/msp.2020.2975749)]

35. Krishnamurthy S, Ks K, Dovgan E, Luštrek M, Gradišek Piletič B, Srinivasan K, et al. Machine learning prediction models for chronic kidney disease using national health insurance claim data in Taiwan. *Healthcare (Basel)* 2021 May 07;9(5):546 [FREE Full text] [doi: [10.3390/healthcare9050546](https://doi.org/10.3390/healthcare9050546)] [Medline: [34067129](https://pubmed.ncbi.nlm.nih.gov/34067129/)]
36. Nemati M, Zhang H, Sloma M, Bekbolsynov D, Wang H, Stepkowski S, et al. Predicting kidney transplant survival using multiple feature representations for HLAs. *Artif Intell Med Conf Artif Intell Med (2005-)* 2021 Jun;12721:51-60 [FREE Full text] [Medline: [34179894](https://pubmed.ncbi.nlm.nih.gov/34179894/)]
37. Nelson RG, Grams ME, Ballew SH, Sang Y, Azizi F, Chadban SJ, CKD Prognosis Consortium. Development of risk prediction equations for incident chronic kidney disease. *JAMA* 2019 Dec 03;322(21):2104-2114 [FREE Full text] [doi: [10.1001/jama.2019.17379](https://doi.org/10.1001/jama.2019.17379)] [Medline: [31703124](https://pubmed.ncbi.nlm.nih.gov/31703124/)]
38. Hsieh C, Su C, Shao S, Sung S, Lin S, Kao Yang YH, et al. Taiwan's National Health Insurance Research Database: past and future. *Clin Epidemiol* 2019;11:349-358 [FREE Full text] [doi: [10.2147/CLEPS196293](https://doi.org/10.2147/CLEPS196293)] [Medline: [31118821](https://pubmed.ncbi.nlm.nih.gov/31118821/)]
39. Matsushita K, Ballew SH, Astor BC, Jong PED, Gansevoort RT, Hemmelgarn BR, Chronic Kidney Disease Prognosis Consortium. Cohort profile: the chronic kidney disease prognosis consortium. *Int J Epidemiol* 2013 Dec;42(6):1660-1668. [doi: [10.1093/ije/dys173](https://doi.org/10.1093/ije/dys173)] [Medline: [23243116](https://pubmed.ncbi.nlm.nih.gov/23243116/)]
40. Gerke S, Minssen T, Cohen G. Ethical and legal challenges of artificial intelligence-driven healthcare. *Artificial Intelligence Healthcare* 2020:295. [doi: [10.1016/b978-0-12-818438-7.00012-5](https://doi.org/10.1016/b978-0-12-818438-7.00012-5)]
41. Wong A, Otlis E, Donnelly JP, Krumm A, McCullough J, DeTroyer-Coolley O, et al. External validation of a widely implemented proprietary sepsis prediction model in hospitalized patients. *JAMA Intern Med* 2021 Aug 01;181(8):1065-1070 [FREE Full text] [doi: [10.1001/jamainternmed.2021.2626](https://doi.org/10.1001/jamainternmed.2021.2626)] [Medline: [34152373](https://pubmed.ncbi.nlm.nih.gov/34152373/)]
42. Mahidol launches medical, engineering dual degree. *Bangkok Post*. URL: <https://www.bangkokpost.com/thailand/general/1774649/mahidol-launches-medical-engineering-dual-degree> [accessed 2022-06-24]
43. Dual Medicine-Engineering degree. University of Galway. URL: <https://www.nuigalway.ie/physicianeerdegree/> [accessed 2022-06-24]
44. MD+master of engineering dual degree. Duke Biomedical Engineering. URL: <https://bme.duke.edu/masters/degrees/md-meng> [accessed 2022-06-24]
45. Panch T, Duralde E, Mattie H, Kotecha G, Celi LA, Wright M, et al. A distributed approach to the regulation of clinical AI. *PLOS Digit Health* 2022 May 26;1(5):e0000040. [doi: [10.1371/journal.pdig.0000040](https://doi.org/10.1371/journal.pdig.0000040)]

Abbreviations

AI: artificial intelligence

CT: computed tomography

FDA: Food and Drug Administration

GMLP: Good Machine-Learning Practice

ML: machine learning

NHS: National Health Service

PROBAST-AI: Prediction Model Risk of Bias Assessment Tool–Artificial Intelligence

SPS: Software as a Medical Device Pre-Specifications

TPLC: Total Product Lifecycle

TRIPOD-AI: Transparent Reporting of a Multivariable Prediction Model for Individual

Edited by K El Emam; submitted 24.09.22; peer-reviewed by B Mesko, M Fernandes; comments to author 19.10.22; revised version received 11.12.22; accepted 28.12.22; published 16.01.23.

Please cite as:

Mashar M, Chawla S, Chen F, Lubwama B, Patel K, Kelshiker MA, Bachtiger P, Peters NS

Artificial Intelligence Algorithms in Health Care: Is the Current Food and Drug Administration Regulation Sufficient?

JMIR AI 2023;2:e42940

URL: <https://ai.jmir.org/2023/1/e42940>

doi: [10.2196/42940](https://doi.org/10.2196/42940)

PMID:

©Meghavi Mashar, Shreya Chawla, Fangyue Chen, Baker Lubwama, Kyle Patel, Mihir A Kelshiker, Patrik Bachtiger, Nicholas S Peters. Originally published in *JMIR AI* (<https://ai.jmir.org>), 16.01.2023. This is an open-access article distributed under the terms of the Creative Commons Attribution License (<https://creativecommons.org/licenses/by/4.0/>), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in *JMIR AI*, is properly cited. The

complete bibliographic information, a link to the original publication on <https://www.ai.jmir.org/>, as well as this copyright and license information must be included.

Original Paper

Deep Learning Transformer Models for Building a Comprehensive and Real-time Trauma Observatory: Development and Validation Study

Gabrielle Chenais^{1*}, MMid, MScPH, MPHDS; Cédric Gil-Jardiné^{1,2*}, MD, PhD; Hélène Touchais^{1*}, MCPM; Marta Avalos Fernandez^{1,3*}, PhD; Benjamin Contrand^{1*}, MSc; Eric Tellier^{1,2*}, MD; Xavier Combes^{2*}, MD; Loick Bourdois^{1*}, MSc; Philippe Revel^{2*}, MD; Emmanuel Lagarde^{1*}, PhD

¹Unit 1219, Bordeaux Public Health Center, Institut National de la Santé et de la Recherche Médicale, Bordeaux, France

²Emergency Department, Bordeaux University Hospital, Bordeaux, France

³Statistics in Systems Biology and Translational Medicine Team, University of Bordeaux, Institut National de Recherche en Sciences et Technologies du Numérique, Talence, France

* all authors contributed equally

Corresponding Author:

Gabrielle Chenais, MMid, MScPH, MPHDS

Unit 1219

Bordeaux Public Health Center

Institut National de la Santé et de la Recherche Médicale

146 rue Léo Saignat

Bordeaux, 33000

France

Phone: 33 33 05 57 57 15

Email: gabrielle.chenais@u-bordeaux.fr

Abstract

Background: Public health surveillance relies on the collection of data, often in near-real time. Recent advances in natural language processing make it possible to envisage an automated system for extracting information from electronic health records.

Objective: To study the feasibility of setting up a national trauma observatory in France, we compared the performance of several automatic language processing methods in a multiclass classification task of unstructured clinical notes.

Methods: A total of 69,110 free-text clinical notes related to visits to the emergency departments of the University Hospital of Bordeaux, France, between 2012 and 2019 were manually annotated. Among these clinical notes, 32.5% (22,481/69,110) were traumas. We trained 4 transformer models (deep learning models that encompass attention mechanism) and compared them with the term frequency–inverse document frequency associated with the support vector machine method.

Results: The transformer models consistently performed better than the term frequency–inverse document frequency and a support vector machine. Among the transformers, the GPTanam model pretrained with a French corpus with an additional autosupervised learning step on 306,368 unlabeled clinical notes showed the best performance with a micro F_1 -score of 0.969.

Conclusions: The transformers proved efficient at the multiclass classification of narrative and medical data. Further steps for improvement should focus on the expansion of abbreviations and multioutput multiclass classification.

(JMIR AI 2023;2:e40843) doi:[10.2196/40843](https://doi.org/10.2196/40843)

KEYWORDS

deep learning; public health; trauma; emergencies; natural language processing; transformers

Introduction

Background

The objective of public health surveillance is to describe a health event in the population to estimate its burden based on its characteristics (incidence, prevalence, survival, and mortality) and evolution. This surveillance contributes to the definition, implementation, monitoring, and evaluation of public health policies. It must also be able to alert to the emergence of new threats to public health (infectious or environmental in origin and natural or terrorist) and monitor and evaluate the impact of known and expected events (seasonal epidemics) or unexpected events (industrial disasters and extreme weather events) on the health of the population. Public health surveillance relies on the collection of data, often in near real time.

The SurSaUD (Surveillance Sanitaire des Urgences et des Décès) syndromic surveillance system was created for the purpose of public health surveillance in France in 2004 by Santé Publique France, the French National Public Health Agency. The SurSaUD system collects daily data from 4 sources: emergency departments (EDs; OSCOUR ED network) [1], emergency general practitioners (SOS Médecins network), crude mortality (civil status data), and electronic death certification including causes of death [2]. Since its inception, the OSCOUR network has recorded >130 million ED visits. Data are collected by the direct extraction of information from patients' electronic health records (EHRs) in a common format for the entire territory and transmitted to Santé Publique France via the OSCOUR network. Owing to the coding of the main diagnosis (International Classification of Diseases [ICD] 10th Revision codes) and progressive improvement of data quality [3], the network can establish real-time monitoring of public health events such as epidemics of influenza, gastroenteritis, or bronchiolitis [4-7]. This is one of the tools currently used to monitor responses to the COVID-19 epidemic in France.

Approximately one-third of ED visits in France are the result of trauma [8]. Trauma is a major cause of mortality and morbidity worldwide [7]. In 2017, trauma and injury accounted for 7.01% (range 6.75%-7.33%) of the deaths in France [9]. Unfortunately, little information is available regarding trauma; although we can know the nature of the main injury, nothing is known about the mechanism (road accident, assault, suicide, etc). However, this information is available in the EHR but in a free-text form. In fact, each time a patient visits the ED, the nurse in charge of reception and orientation and the physician in charge of the first consultation enter a text called clinical note, which describes the reasons for the patient's visit and the circumstances in which the symptoms occurred. To add the trauma mechanism to the data collected by the OSCOUR network, a manual classification by health professionals would be time consuming and require multiple resources. Given the nature of the data (free text, unstructured, and containing abbreviations) to be processed and the objective (classification), artificial intelligence with deep learning, particularly automatic language processing, seems to be indicated.

Natural language analysis has seen a recent breakthrough with the introduction of deep learning, in particular, the transformer

architecture. Introduced in 2017 by Google and proposed in the article "Attention is All You Need" by Vaswani et al [10], transformers have an architecture that allows the implementation of a mechanism for processing the sequence of tokens (a token is an instance of a sequence of characters in a particular document that are grouped together as a semantic unit useful for language processing) that form a sentence in a self-attentive manner, that is, relating each of these tokens to each of the others in the sentence. They have the particularity of being able to be pretrained on a corpus of text, which can be very large because it does not require a coding stage. This phase leads to a generative model that is capable, for example, of constructing artificial text by iteration. The Bidirectional Encoder Representations from Transformers (BERT) are one of these transformer-type models pretrained on large corpora of text [11]. The BERT model is a bidirectional transformer composed of only encoder blocks. The particularity of BERT model is that it learns information from both the right and left sides of a token's context during the pretraining and training phases. BERT is composed of a stack of 12 identical layers. Each layer consists of 2 sublayers. The first is a multihead self-attention mechanism, and the second is a simple, position-wise fully connected feed-forward network. In other words, the text encoder converts text into a numeric representation. On many tasks, including text classification, its performance is systematically superior to that of the convolutional and autoregressive models used until then [11].

French derivatives of the BERT model such as FlauBERT [12] and CamemBERT [13] have been trained on very large and diverse French corpora. FlauBERT is a French BERT trained on a very large and heterogeneous French corpus. Models of different sizes were trained using the Jean Zay supercomputer of the Centre National de la Recherche Scientifique; there are 3 sizes: small (54 million parameters), base cased (138 million parameters) and uncased (137 million parameters), as well as large (373 million parameters). CamemBERT is based on RoBERTa [14], which is an evolution of BERT in several aspects, including the use of the masked language model as the sole pretraining objective. Similar to FlauBERT, CamemBERT is available in different sizes: base (110 million parameters) and large (335 million parameters); moreover, it can be trained on different training corpora such as OSCAR (either 138 GB or 4 GB of text) [15], CCNET (either 135 GB or 4 GB) [16], or French Wikipedia (4 GB).

One of the most interesting examples of transformer architecture is Generative Pretrained Transformer-2 (GPT-2), released by OpenAI in 2019. GPT-2 is a large transformer-based model composed solely of decoder blocks, with 1.5 billion parameters on its extra-large version, and trained on a data set of 8 million web pages to predict the next word from the previous words [17]. A total of 3 other sizes of GPT-2 were released before the largest: 124 (small), 355 (medium), and 774 (large) million parameters. This model's ability to generate text attracted the attention of the community quickly because of the difficulty in distinguishing the produced artificial texts from the texts written by humans, suggesting that some of the meaning present in natural language was embedded. Moreover, beyond its ability to generate coherent texts, GPT-2 can perform other tasks such

as answering questions or classifying documents. As with BERT, the conservation of several self-attention block weights from a pretrained model is sufficient to transfer contextual representations into another data set. The training of the GPT-2 model is thus carried out in 2 distinct phases. The first phase of self-supervised generative pretraining consists of the reading of a corpus of texts. This leads to the ability to generate texts automatically. The second supervised training phase consists of resuming the learning process in a corpus of annotated texts to create a system capable of performing specific tasks (eg, classification). BelGPT2 is a Belgian small GPT-2 pretrained on a French corpus of 60 GB (Common Crawl, Project Gutenberg, Wikipedia, EuroPARL, etc) that was released at the end of 2020 [18].

Related Work

Extracting mechanisms and types of traumas are a matter of multiclass classification. Multiclass classification of French medical data involves a wide variety of techniques. For example, for the 2018 Conference and Labs of the Evaluation Forum eHealth task 1 challenge [19], the objective of which was to extract ICD 10th Revision codes from the death certificates provided by the Centre for Epidemiology of Medical Causes of Death, Cossin et al [20] tested an approach based on ontologies, whereas Flicoteaux et al [21] proposed an approach using a probabilistic convolutional neural network (CNN), and Ive et al [22] resorted to the association of a recurrent neural network with a CNN. By contrast, Metzger et al classified free-text clinical notes from ED related to suicide attempts using random forest and naive Bayes-type algorithms [23]. Recent studies have shown the effectiveness of transformers in classification tasks for EHR free-text data such as ICD coding [24,25], phenotyping [26], and readmission prediction [27]. Therefore, within the framework of the TARPON (Traitement Automatique des Résumés de Passage aux urgences dans le but de créer un Observatoire National) project, which aims to demonstrate the feasibility of setting up a national observatory of trauma, we propose here to compare the performances of several transformer models in the classification of ED visits for trauma based on clinical notes from the adult ED of the Bordeaux University Hospital. We compared the transformers FlauBERT, CamemBERT, BelGPT2, and a French GPT-2 model pretrained on a domain-specific corpus called GPTanam with term frequency-inverse document frequency (TF-IDF)/support vector machine (SVM), which was used as a baseline model. To the best of our knowledge, no previous performance evaluation of multiple transformers' classification application has been conducted on complex and unstructured clinical data from ED combining common French language, medical data, and jargon.

Methods

Medical Ethics Regulations and General Data Protection Regulation

This study was authorized by the Bordeaux University Hospital Ethical Board under number GP-CE2021-21. A data management plan was created and reviewed by the privacy security board to meet the institutional and national requirements in France for General Data Protection Regulation compliance.

Database

Clinical notes were extracted from the EHR of the adult ED stored in the information system of the University Hospital of Bordeaux, France. They correspond to 375,478 medical records of visits to the adult ED of Bordeaux Hospital from 2012 to 2020. The variables available were age, sex, date and time of the visit, the clinical note generated by the physicians or interns, and the clinical note written by the triage nurses.

Labeling Strategy

In total, 69,110 clinical notes were randomly extracted for manual annotation. Our coding team consisted of trauma epidemiologists, emergency physicians, emergency nurses, research assistants, and biostatisticians, amounting to a total of 16 coders. The annotation phase lasted 5 months. For each clinical note, a code describing the content of the text was assigned. The annotation grid used for coding was developed for the needs of the project. The code associated with each clinical note consisted of 9 fields. The fields were as follows: "First visit (to the emergency department for this reason)," "Location (of the trauma)," "Activity (performed during the trauma)," "Type of Sport (practiced during the trauma)," "Subject under the influence," "Notion of pre-traumatic discomfort," "MVA (Motor Vehicle Accident)-Secondary Prevention Elements," "MVA-Antagonist," and "Type of trauma or Mode of travel for the MVA." As the objective was to classify the types of trauma, we mainly used the data of the field "Type of trauma or Mode of movement for the MVA." As the distribution of the fields was unbalanced, we created a composite variable containing 8 mutually exclusive classes to have a larger number of clinical notes per class. Therefore, we grouped certain types of traumas (ie, "Fall," which included "Fall from own height," "Fall from a given height," and "Fall on stairs"). The composite variable included the following classes or labels: "Accident of exposure to body fluids (blood exposure accident, unprotected sex at risk)," "Assault," "Motor Vehicle Accident (MVA)," "Foreign body in eyes," "Fall (except sports)," "Sports accident," "Intentional Injury," and "Other trauma" as shown in [Multimedia Appendix 1](#). The interannotator agreement was assessed with a random sample of 1000 clinical notes labeled by 2 annotators, leading to a Cohen κ score [28] of 0.84.

A sensitivity analysis was performed to study the impact of potentially ambiguous content on classification. Therefore, the test sample was reread by an expert. Potentially ambiguous content in terms of classification is defined here as the accumulation of several mechanisms or types of traumas or a major difficulty in assigning a label to a clinical note given its text.

Corpus Statistics

In total, 22,481 manually labeled clinical notes from the Bordeaux University Hospital were included in the study. One-third (22,481/69,110) of the total annotated clinical notes were labeled as visit to the ED resulting from a trauma. The average number of sentences in the corpus was 3.25 (SD 2.56; range 1-63). The average length of clinical notes was 58 (SD 38) words, with a minimum of 1 word (eg, "Accident d'exposition au sang") and a maximum of 630 words. The

number of unique unigrams, bigrams, and trigrams were 70,99, 395,827, and 777,459, respectively.

Models and Experiment Settings

The models selected for comparison and freely available as open-source content were a traditional machine learning model (baseline model) with TF-IDF/SVM couple as well as 3 transformer-type models pretrained on French corpora: CamemBERT [13], FlauBERT [12], and BelGPT2 [18]. We then chose the best performing model and applied a supplementary step of self-supervised training with the remaining 306,368 unlabeled clinical notes. This model is called here as GPTanam. Table 1 lists the size and configuration of each transformer model.

For TF-IDF, tokenization was performed using the National Language Toolkit package (version 3.6.6; NLTK) [29], and linear support vector classifier was applied using scikit-learn (version 0.24.1) [30]. The most frequent words (eg, “that,” “he,” and “the”) were removed. Tokenization was performed using SentencePiece [31] for CamemBERT, Byte-Pair Encoding for

FlauBERT, and a byte-level Byte-Pair Encoding for both GPT-2 models [32]. The data were cleaned using regular expressions with the `re` package in Python (version 3.7). Unicode normalization was performed in the 8-bit Universal Character Set Transformation Format. The linear support vector classifier parameters were as follows: tolerance= 1×10^{-5} , penalty=l2, loss=squared hinge, dual optimization=true, C=1.0, multiclass strategy=one versus rest, verbose=0, and a maximum of 1000 iterations. For all 3 transformers, the optimizer was AdamW, with an epsilon of 1×10^{-8} , and the maximum length was 512. GPTanam had training and evaluation batch sizes of 5 and a learning rate of 2×10^{-5} . For FlauBERT and CamemBERT, the batch size was 16 for training and 20 for evaluation, and the learning rate was 5×10^{-5} . The models were trained using the Hugging Face library under the Pytorch framework on our workstation with a single Titan RTX (Nvidia) graphics processing unit with 24 GB of video RAM. Performance analysis was done using scikit-learn and imbalance-learn (version 0.9.1).

Table 1. Transformer models' sizes and configurations.

Model	Layers	Attention heads	Embedding dimension	Parameters (millions)	Pretraining corpus size (GB)
CamemBERT-base-CCNET ^a	12	12	768	110	135
FlauBERT-base-cased	12	12	768	138	71
BelGPT2	12	12	768	117	57.9
GPTanam	12	12	768	117	58.6

^aCCNET: criss-cross attention for semantic segmentation.

Self-supervised Learning and Fine-tuning Phase

Considering the GPTanam model, the first step comprising self-supervised learning was performed with 306,368 clinical notes with 1 epoch [33]. For all the models, a random sample of 80.80% (18,166/22,481) of the clinical notes labeled as trauma was dedicated to supervised learning. This data set was divided into a training sample (14,532/18,166, 79.99%) and a validation sample (3634/18,166, 20%) in an 80/20 ratio. We trained each model 9 times with different seeds on 7 epochs for CamemBERT and FlauBERT and 5 epochs for BelGPT2 and GPTanam. To obtain a single prediction for the 9 different executions of the chosen epoch (based on the maximum validation micro F_1 -score) for each model, a vote was taken.

Test Phase

The test sample contained 19.19% (4315/22,481 records) of the labeled data set. The second reading of these clinical notes resulted in 10.82% (467/4315) being tagged as clinical notes with potentially complex or ambiguous content in terms of classification. Therefore, the analysis included both the complete test data set (4315/22,481, 20%) and the data set without

complex and ambiguous content (3848/22,481, 17.11%). To obtain the probabilities for each prediction, a softmax activation layer was applied to the 4 transformer models.

Data Sets

The label distribution among the corpus and each training, validation, and test data set are presented in Table 2. The most common type of trauma was the class “Fall” followed by “Other trauma” and “Motor Vehicle Accident.” An example of clinical notes translated from French is shown in Multimedia Appendix 2.

The median age at the time of visit was 37 (IQR 24-58—first and third quartiles) years, and 58.46% (13,143/22,481) of the patients were male. EHRs were introduced in 2012 at the Bordeaux University Hospital, which explains the lower proportion of data for this particular year. In 2019, there was a decrease in ED venues, whereas in 2020, there was a significant increase in ED venues. Table 3 summarizes the characteristics of the train, validation, and test data sets for the study population. The distribution of the variables age, sex, and year of venues at the ED were comparable among the 3 data sets.

Table 2. Label distribution among train, validation, and test data sets.

Type of trauma	Train data set (n=14,532, 64.64%), n (%)	Validation data set (n=3634, 16.16%), n (%)	Test data set (n=4315, 19.19%), n (%)	Total (N=22,481, 100%), n (%)
Accident of exposure to bodily fluids	132 (0.91)	40 (1.1)	41 (1)	213 (0.9)
Assault	1587 (10.92)	393 (10.81)	498 (11.54)	2478 (11.02)
MVA ^a	2028 (13.95)	495 (13.62)	568 (13.16)	3091 (13.75)
Foreign body in eye	642 (4.42)	180 (5)	186 (4.31)	1008 (4.48)
Fall	4778 (32.87)	1162 (31.97)	1554 (36.01)	7494 (33.33)
Sport accident	1311 (9)	341 (9.38)	371 (8.59)	2023 (9)
Intentional injury	341 (2.34)	73 (2)	112 (2.59)	526 (2.33)
Other trauma	3713 (25.55)	950 (26.14)	985 (22.82)	5648 (25.12)

^aMVA: motor vehicle accident.

Table 3. Train, validation, and test data set characteristics.

	Train data set (n=14,532)	Validation data set (n=3634)	Test data set (n=4315)	Total (N=22,481)
Age (years), median (IQR ^a)	37 (24-58)	37 (24-57)	37 (24-58)	37 (24-58)
Sex (male), n (%)	8486 (58.39)	2181 (60.01)	2476 (57.38)	13,143 (58.46)
Year of ED^b venue, n (%)				
2012	218 (1.5)	52 (1.43)	66 (1.52)	336 (1.49)
2013	1389 (12.2)	359 (12.4)	418 (12.3)	2166 (12.2)
2014	1444 (12.6)	385 (13.3)	386 (11.3)	2215 (12.3)
2015	1502 (13.1)	326 (11.2)	425 (12.5)	2253 (12.6)
2016	1419 (12.4)	365 (12.6)	426 (12.6)	2210 (12.3)
2017	1493 (13.1)	370 (12.8)	461 (13.5)	2324 (12.9)
2018	1425 (12.5)	405 (13.9)	474 (13.9)	2304 (13.5)
2019	690 (6)	175 (6)	218 (6.4)	1083 (6.2)
2020	1856 (16.2)	468 (16.1)	532 (15.6)	2856 (16)
Missing values	3118 (27.3)	737 (25.4)	899 (26.4)	4724 (20.9)

^aIQR: first and fourth quartiles are given.

^bED: emergency department.

Performance Criteria

The measures chosen were macro-average precision and micro F_1 -score, which, in the multiclass framework, are equal to accuracy. For the following equations, n is the number of samples (clinical notes), TP is true positive, FP is false positive, and FN is false negative.

Macro-Average Precision

Precision expresses the proportion of units a model classifies as positive that are actually positive. In other words, precision indicates how much one can trust the model when it predicts that a record is classified in a given class. In the case of multiclass classification, the macro-average precision over all i classes can be evaluated by macro-averaging, wherein the precision over each i class is first calculated and then the precisions over all n classes are averaged. There is no relation to class size, as classes of different sizes are also weighted in

the numerator. This implies that the effect of larger classes is as important as that of smaller ones. Therefore, each clinical note is equally important with this measure [34].



Micro F_1 -Score

F_1 -score is defined as the harmonic mean of precision and recall in binary class problem. To extend the F_1 measure to multiclass, 2 types of average, microaverage and macro-average, are commonly used. In microaveraging, the F_1 measure is computed globally over all class decisions, with precision and recall being obtained by summing over all individual decisions. The microaveraged F_1 measure gives equal weight to each clinical note and is, therefore, considered as an average over all the clinical note or category pairs [35].



Data Security

Identifying information was found in the data set. Therefore, we deidentified all clinical notes using named entity recognition with FlauBERT. Data processing and computing were conducted within the facilities of the ED of the University Hospital of Bordeaux, which have received regulatory clearance to host and exploit databases with personal and medical data. All the patients from whom information was retrieved were aged ≥ 15 years.

Error Analysis

An error analysis was performed with unigrams and bigrams for the best performing model. All misclassified clinical notes were read by an expert to determine whether the human annotation labels were appropriate.

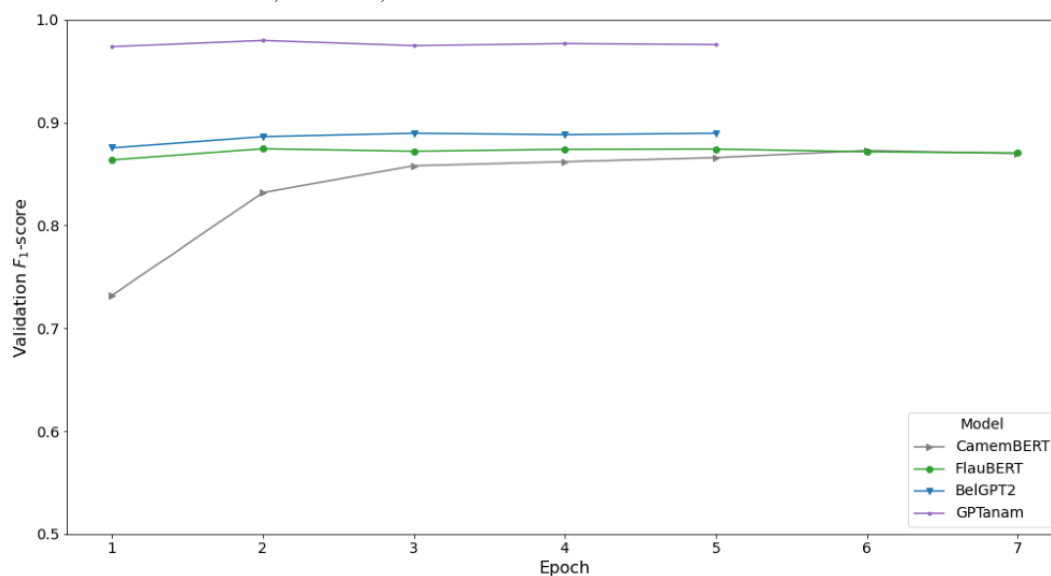
Results

Fine-tuning the Performance of Models

Unlike statistical methods such as TF-IDF, the supervised fine-tuning of transformer models is time consuming and can

be greatly accelerated by the use of graphics processing units. The self-supervised fine-tuning step for the GPTanam model required approximately 12 hours. At that point, GPTanam could generate artificial clinical notes, as seen in [Multimedia Appendix 3](#), that could not be easily differentiated from the original ones. One epoch of supervised fine-tuning required 15, 16, 19, and 18 minutes for CamemBERT, FlauBERT, BelGPT2, and GPTanam, respectively. When looking deeper into each transformer model's F_1 -scores on the validation data set, [Figure 1](#) shows that CamemBERT reached its maximum F_1 -score (0.873) at epoch 6, FlauBERT achieved an F_1 -score of 0.874 at epoch 5, BelGPT2 reached its peak (0.890) faster at epoch 3, and GPTanam reached 0.980 at epoch 2. Moreover, GPTanam's F_1 -score on the validation data set was the highest among the 4 transformer models. We conjecture that a self-supervised step on a domain-specific corpus for GPTanam contributed to the learning of the semantic representations, which resulted in a faster convergence in the learning of the classification task.

Figure 1. F_1 -score curves for CamemBERT, FlauBERT, BelGPT2 and GPTanam on the validation dataset.



Performance of Models

The average macro precision and micro F_1 -scores were systematically higher for the transformers than for the TF-IDF/SVM couple on the complete test data set, as shown in [Table 4](#). Among the transformers, GPTanam achieved an average micro F_1 -score of 0.969, outperforming CamemBERT, FlauBERT, and BelGPT2, for which average F_1 -scores were 0.878, 0.873, and 0.887, respectively. The macro-average precision was higher than the F_1 -score in almost all cases, except for TF-IDF/SVM, for which the macro precision was lower than the micro F_1 -score (macro precision=0.860 and micro F_1 -score=0.864).

The distribution of n clinical notes per class was not balanced, and the micro F_1 scores were, in all cases, lower in the classes where n was lower. Concerning the micro F_1 -score of the different classes, GPTanam had higher scores than the other transformers and TF-IDF. The performance of GPTanam was high for all classes except for intentional injuries; we assumed that this might be associated with the semantic heterogeneity and variety of the class. Indeed, this class encompassed self-harm (self-mutilation, punching due to rage, and self-stabbing) and suicide attempts (shooting, alcohol or drug poisoning, and car crashing), with few examples per injury. By contrast, classes such as motor vehicle accident (MVA) and fall have semantic consistency with a larger number of examples. The confusion matrix is shown in [Multimedia Appendix 4](#). An

error analysis of the intentional injury class, as well as the other classes, is provided in the next section.

Table 4. Micro F1-scores for all classes and models with microaverage F1-scores and macro-average precision on the complete test data set.

Type of trauma	Test data set (n=4315), n (%)	Micro F_1 -scores				
		TF-IDF ^a /SVM ^b	CamemBERT	FlauBERT	BelGPT2	GPTanam
Accident of exposure to bodily fluids	41 (1)	0.83	0.84	0.84	0.83	<i>0.91</i> ^c
Assault	498 (11.54)	0.9	0.91	0.92	0.91	<i>0.96</i>
MVA ^d	568 (13.16)	0.91	0.90	0.91	0.91	<i>0.97</i>
Foreign body in eye	186 (4.3)	0.79	0.84	0.82	0.82	<i>0.97</i>
Fall	1554 (36.01)	0.9	0.92	0.91	0.92	<i>0.98</i>
Sport accident	371 (8.6)	0.82	0.83	0.83	0.85	<i>0.94</i>
Intentional injury	112 (2.6)	0.75	0.76	0.73	0.77	<i>0.84</i>
Other trauma	985 (22.8)	0.8	0.83	0.82	0.85	<i>0.98</i>
Micro F_1 -score	N/A ^e	0.864	0.878	0.873	0.887	<i>0.969</i>
Macro precision	N/A	0.860	0.880	0.880	0.89	<i>0.970</i>

^aTF-IDF: term frequency–inverse document frequency.

^bSVM: support vector machine.

^cThe best F_1 -scores are in italic.

^dMVA: motor vehicle accident.

^eN/A: not applicable.

Error Analysis

The error analysis results are presented in [Textbox 1](#).

Removing complex and ambiguous clinical notes were associated with an increase of performance for all the models; the average gain of F1-scores was 0.04 for TF-IDF/SVM, CamemBERT, FlauBERT, and BelGPT2. The average gain of the micro F1-score was 0.01 for GPTanam, which seems to be more robust in classifying complex and ambiguous content.

The difference in performance when potentially complex and ambiguous content was considered was greater for TF-IDF/SVM, CamemBERT, FlauBERT, and BelGPT2 than for GPTanam, especially with the classes MVA and Sport

Accident, where the average gain of the micro F1-score per class was 0.07, as shown in [Figure 2](#). Performance for the class “Accident of exposure to bodily fluids” did not improve for TF-IDF/SVM, CamemBERT, and FlauBERT when complex and ambiguous content was removed from the test data set. The performance of GPTanam did not improve for the classes “Foreign body on the eye” and “Other trauma,” but the F1-scores were already very high for these classes—0.97 and 0.98, respectively. Performance was slightly improved for “Assault,” “Fall,” “MVA,” “Sport Accident,” and “Other trauma” when potentially complex and ambiguous content was removed from the test data set for all the models as seen in [Multimedia Appendix 5](#) and the confusion matrix in [Multimedia Appendix 6](#).

Textbox 1. Error analysis results.**Accident of exposure to bodily fluids**

The bigram analysis showed that the keywords “contact blood” were absent in the top 10 bigrams in the incorrectly classified clinical notes, whereas the unigrams analysis showed that “HIV” is the ninth unigram (after “aes,” “blood,” “needle,” “source,” “intercourse,” “dakin,” “work,” and “sexual”).

Assault

Regarding the class “Assault,” the top 3 bigrams were “physical assault,” “declare having,” and “punch” (*coup poing* in French) for the correctly classified clinical notes, whereas “left hand,” “hand trauma,” and “mechanical fall” were the most frequent bigrams. The verification of the 18 clinical notes manually annotated as “Assault” showed that for 11 (61%) of them, the label predicted by the model was correct (n=1, 9% fall; n=8, 73% self-harms; n=1, 9% motor vehicle accident [MVA]; and n=1, 9% sport accident paintball).

MVA

The acronym “mva” (n=700, 26%) was the most represented unigram in the correctly classified corpus, whereas “pain” was the most represented unigram in the clinical notes classified as not MVA. When analyzing the 6 incorrectly classified clinical notes, 3 (50%) of them were wrongly labeled as they were in fact referring to an assault, a fall, and a basketball accident. The 3 (50%) remaining clinical cases contained 2 types of traumas such as falling on the street.

Foreign body in the eye

The unigram analysis for this class showed that the unigrams “eye” and “the eye” were the most represented (n=140, 13%), whereas “left” and “hear” were the top 2 unigrams in the clinical notes classified as not being “foreign body in the eye.” In fact, one of these clinical notes was related to a foreign body in the heart, and 2 others were assault without mention of eye trauma.

Fall

The top 3 bigrams for the correctly classified clinical notes were “mechanical fall,” “loss of consciousness,” and “cranial trauma” and “right ankle,” “ankle trauma,” “left ankle” for the incorrectly classified ones. In total, 21 of the incorrectly classified clinical notes encompassed a double mechanism of trauma: 1 (5%) sport accident, 16 (76%) MVAs, and 4 (19%) assaults involving a fall were present. A total of 9 notes mentioned back pain, ankle and knee twists, pain while getting off of a truck, or a patient found at the bottom of the stairs without mention of falling.

Intentional injury

The most frequent unigrams and bigrams were different between the correctly and incorrectly classified clinical notes. The most represented unigrams and bigrams were, respectively, “imv” (“voluntary drug intoxication” in French) and “suicide attempt” in the correctly classified corpus of clinical notes, whereas “hand” and “punch given” were the most common in the incorrectly classified notes. Indeed, the model classified 10 clinical notes as assault, whereas these clinical notes were related to a patient having punched something or himself.

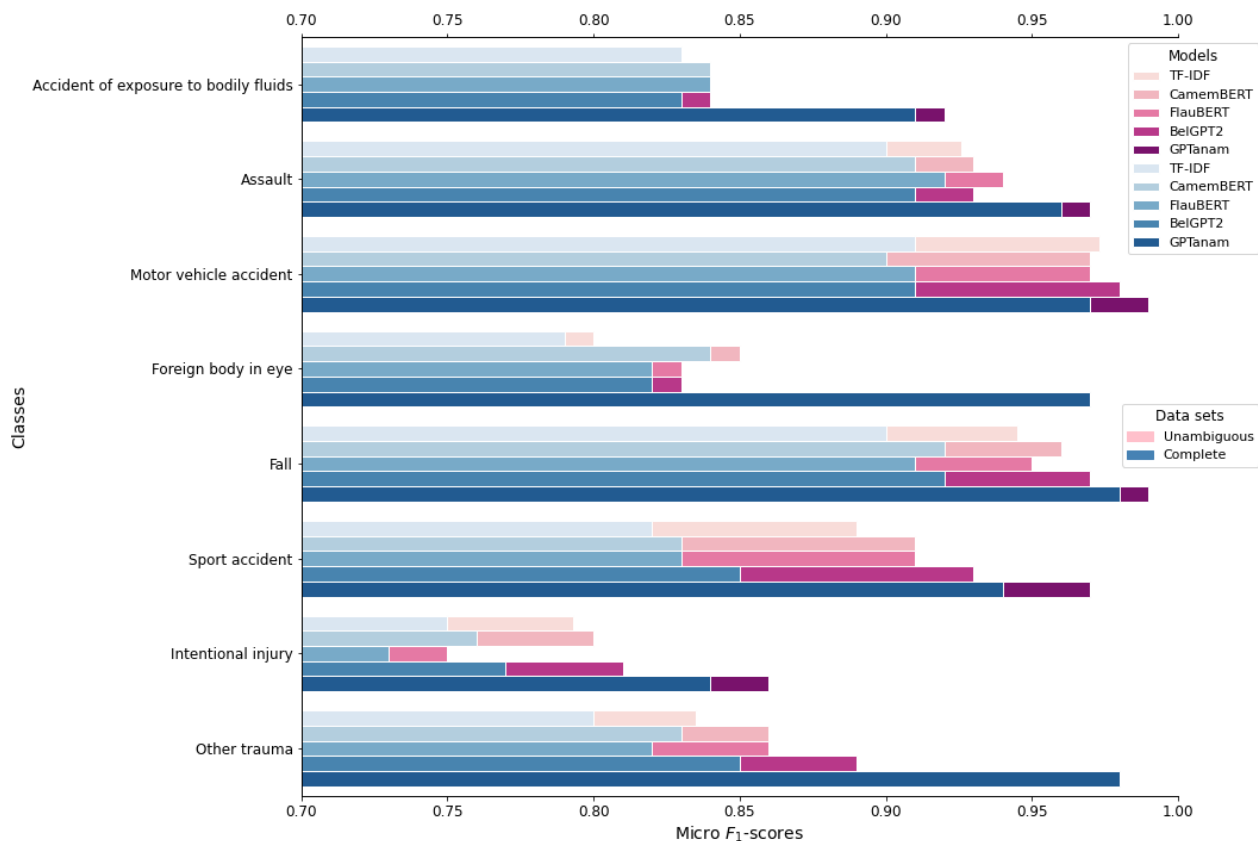
Sport

In the correctly classified clinical notes, the most frequent unigrams were “pain,” “left,” and “trauma” and the most frequent bigrams were “right ankle,” “functional impotence,” and “left knee.” In the incorrectly classified notes, the most frequent unigrams and bigrams were, respectively, “fall,” “trauma,” and “bike” and “bike fall,” “right knee,” and “knee pain.” A total of 13 falls occurred while biking (the notes did not mention the place) and were classified as MVA. Five incorrectly classified notes were eye trauma while practicing sports.

Removing complex and ambiguous clinical notes were associated with an increase of performance for all the models; the average gain of F_1 -scores was 0.04 for TF-IDF/SVM, CamemBERT, FlauBERT, and BelGPT2. The average gain of the micro F_1 -score was 0.01 for GPTanam, which seems to be more robust in classifying complex and ambiguous content.

The difference in performance when potentially complex and ambiguous content was considered was greater for TF-IDF/SVM, CamemBERT, FlauBERT, and BelGPT2 than for GPTanam, especially with the classes MVA and Sport Accident, where the average gain of the micro F_1 -score per class was 0.07, as shown in [Figure 2](#). Performance for the class “Accident of exposure to bodily fluids” did not improve for TF-IDF/SVM, CamemBERT, and FlauBERT when complex and ambiguous content was removed from the test data set. The performance of GPTanam did not improve for the classes “Foreign body on the eye” and “Other trauma,” but the F_1 -scores were already very high for these classes—0.97 and 0.98, respectively. Performance was slightly improved for “Assault,” “Fall,” “MVA,” “Sport Accident,” and “Other trauma” when potentially complex and ambiguous content was removed from the test data set for all the models as seen in [Multimedia Appendix 5](#) and the confusion matrix in [Multimedia Appendix 6](#).

Figure 2. Plot of micro F1-scores of all models for each class for both the complete test data set (blue bars) and the test data set without potentially ambiguous content as regard to its classification (pink bars). TF-IDF: term frequency–inverse document frequency.



Discussion

Transformers: A New State of the Art

The transformers showed interesting results when applied to free-text data from the ED of the Bordeaux University Hospital; a GPT-2 model with a French tokenizer and a self-supervised training step on a domain-specific corpus in addition to a large French corpus reached an average micro F_1 -score of 0.969. This model showed better performance than TF-IDF/SVM and the other transformer models on average metrics and for all classes. In 2018, when reviewing deep learning algorithms for clinical natural language processing, the study by Wu et al projected the rise in the popularity of transformer models [36]. However, some studies showed that traditional approaches, when tailored to the specific language and structure of the text inherent to the classification task, can achieve or exceed the performance of more recent ones based on contextual embeddings such as BERT [37]. Further study could involve comparing our model's performance with that of bidirectional long short-term memory with pretrained embeddings such as Word2Vec or transformer embeddings and CNN.

Self-supervised Training on Domain-Specific Corpus and Tokenizer

The decision to use pretrained models on French corpora with a French tokenizer has probably contributed to the global performance of the chosen transformer models. General language transformer models pretrained on a cross-domain text corpus in a given language have recently flourished. BelGPT2

was the first GPT-2 model fine-tuned on a French heterogeneous corpus (CommonCrawl, French Wikipedia, and EuroParl) released on the Hugging Face platform. The self-supervised training of transformers in a specific domain can improve the performance of tasks such as classification [38], text generation [39], and predicting hospital readmission [40]. Despite many experiments using BERT, GPT-2 has not been studied as extensively as BERT yet. Our team showed that the amount of data required to achieve a given level of performance (area under the curve >0.95) was reduced by a factor of 10 when applying self-supervised training on emergency clinical notes to a binary classification task [41]. Here, we confirmed the benefits of a self-supervised training step on a domain-specific corpus. However, it is questionable whether this approach will be applicable when extending the TARPON project to data from other EDs in France, as each region or ED uses a specific language in addition to the medical language, which uses many abbreviations that can vary locally (eg, assault is written as “brawl” in Bordeaux and “hep” means hepatitis). A possible solution would be to train the model on a corpus resulting from the extraction of ED notes at a national level. Similarly, the treatment of medical concepts and abbreviations remains an area for improvement, as not all EDs use the same abbreviations in the same context. The use of ontologies developed in the field of emergencies could constitute an area for improvement. Transformers have also recently been tested for the identification and replacement of abbreviations, with good results for BERT [42,43]; however, there has not yet been a test on data from a mixture of common language and medical terms in French.

In addition, because the authors who proposed the CamemBERT model did not compare the performance of different models from the OSCAR, CCNET, and Wikipedia data sets in a classification task, a future study could compare the different sets in our database in this regard. While we have only used the basic models of CamemBERT, FlauBERT, and GPT-2, it would be appropriate to test the different sizes of pretraining data sets on a classification task as well as the different sizes of models. Indeed, Martin's [44] team has shown that the standard CamemBERT model (110 million parameters) trained on all 138 GB of OSCAR text does not massively outperform the model trained "only" on the 4 GB sample in morphosyntactic labeling, syntactic parsing, named entity recognition, and natural language inference. One perspective considered is to test different models of French transformers that have been released since CamemBERT, FlauBERT, and BelGPT2 such as Pagnol and BARThez.

Taxonomy

The performance of the models improved when we excluded the clinical notes that we considered the most complex and ambiguous from our test data set. The classification error analysis showed that when clinical notes encompassing 2 mechanisms of trauma (ie, "fall from bike on the street") were removed from the test data set, the models performed better. This expected result shows that since the advent of transformers, the margin of progress in a free-text classification task is nowadays low. This behavior was less important with GPTanam, which seems to have benefited from the self-supervised pretraining phase for reducing classification errors by learning semantic representations beforehand. However, the annotation grid created for the project is partly responsible for some classification errors in the sense that there are areas of semantic overlap between classes. In addition, the coding system used did not allow for the coding of several traumatic mechanisms (eg, a collision between 2 individuals followed by a fall). To be able to account for these situations, a new coding system will be used for the next phases of the project, using the recently released version of trauma classification grid used by the FEDORU (Fédération des Observatoires Régionaux des Urgences) and OSCOUR.

Improving Trauma Public Health Surveillance

The costs of injury and morbidity are immense not only in terms of lost economic opportunities and demands on national health

budgets but also in terms of personal suffering [45]. However, few countries have surveillance systems that generate reliable information on the nature and extent of injuries, especially nonfatal injuries. The traditional view of injuries as "accidents" or random events has resulted in the historical neglect of this area of public health [46]. However, in recent decades, public health officials have been recognizing traumas as preventable events and have been promoting evidence-based interventions for the prevention of traumas worldwide [47]. Many injury interventions are already in place (eg, transportation requirements such as setting speed limits, safe automobile design, seatbelt and other safety restraint use, and use of helmet and other protective equipment) and have achieved significant public health improvements, including the reduction of trauma occurrence [48].

The automatic labeling of ED clinical notes will contribute to an effective real-time public health surveillance system for traumas. Future steps encompass deployment in hospitals' IT departments in Gironde, France, at first, and then at a national scale.

Conclusions

Transformers have shown great effectiveness in a multiclass classification task on complex data encompassing narrative, medical data, and jargon. The choice of this type of architecture in the automatic processing of ED summaries to create a national observatory is relevant. Applying a self-supervised training step on a specific domain corpus has substantially improved the classification performance of a French GPT-2 model. The next labeling strategy within the framework of the TARPON project will be carried out using a standardized trauma classification tool, which will allow a more precise classification of trauma mechanisms owing to a clearer delineation between the different classes (little overlap of semantic fields). The objective is eventually to have a single code for ED summaries, including several variables (eg, place of occurrence, activity during the trauma, and role in a road accident). It is necessary to investigate the possibility of making predictions with a model trained on each variable or using a single model trained on all variables. If the latter method is chosen, a larger model of GPT-2 will probably be required. Furthermore, the expansion of acronyms is under consideration in the automation pipeline.

Acknowledgments

This work was carried out within the framework of the TARPON (Traitement Automatique des Résumés de Passages aux urgences pour un Observatoire National) project led by the Inserm team Injury Epidemiology (project leader E Lagarde) and the emergency department of the Bordeaux University Hospital in collaboration with the Statistics In System biology and Translational Medicine team, managed by Inria and Inserm. This project is the winner of the second call for projects of the Health Data Hub, Grand Défi "Improving medical diagnosis through Artificial Intelligence" and Bpifrance. This study was conducted within the framework of PIA3 (Investment for the Future; project number 17-EURE-0019). The authors would like to thank all the members of the labeling team. The authors would also like to thank the University Hospital of Bordeaux for providing logistical support, which allowed the authors to access and analyze the data needed for the manuscript in such a short period. They are also grateful to Julien Anjoubault, Clarisse Marguinaud, Virginie Cocuelle, Delphine Vauthier, Alexandra Barbe, François Garreau, Quentin Bana, Claire Riou, Pauline Soubelet, and Elisabeth Verbitskaya for their expertise, which allowed proper manual coding for

validation, and to Benjamin Contrand and Marie-Odile Coste for data management and administrative assistance. Bordeaux Population Health Injury Epidemiology Transport Occupation Team activities are supported by the Institut National de la Santé et de la Recherche Médicale, University of Bordeaux, and Ministère de l'Intérieur (Délégation à la Sécurité Routière).

Data Availability

The data set is not available because of patient privacy restrictions. However, the model may be shared with qualified researchers from academic or university institutions upon request via the corresponding author.

Authors' Contributions

EL and GC designed the experiments. GC drafted the paper. HT and GC programmed the design of the experiments. The scripts were checked together by HT and GC. GC designed the data set. CGJ extracted the data set from the database. The paper was revised by all the authors. Guarantor is GC.

Conflicts of Interest

None declared.

Multimedia Appendix 1

Composite variable creation. MVA: motor vehicle accident.

[[PNG File , 461 KB - ai_v2i1e40843_app1.png](#)]

Multimedia Appendix 2

Emergency department electronic health record visualization with clinical note translated in English.

[[PNG File , 254 KB - ai_v2i1e40843_app2.png](#)]

Multimedia Appendix 3

Example of 2 clinical notes artificially generated by GPTanam right after the self-supervised training step with a setup of maximum 40 tokens generated. Clinical notes in French are on the left, and translated notes in English are on the right.

[[PNG File , 41 KB - ai_v2i1e40843_app3.png](#)]

Multimedia Appendix 4

Confusion matrix for the GPTanam model on the complete test data set. Ratio and percentage of correctly classified clinical notes per class are given. MVA: motor vehicle accident.

[[PNG File , 3054 KB - ai_v2i1e40843_app4.png](#)]

Multimedia Appendix 5

Average macro-precision and micro F1-score for each model for the test data set without complex/ambiguous content in clinical notes. MVA: motor vehicle accident; SVM: support vector machine; TD-IDF: term frequency–inverse document frequency;.

[[PNG File , 25 KB - ai_v2i1e40843_app5.png](#)]

Multimedia Appendix 6

Confusion matrix for the GPTanam model on the test data set without complex/ambiguous content in clinical notes. Ratio and percentage of correctly classified clinical notes per class are given. MVA: motor vehicle accident.

[[PNG File , 2995 KB - ai_v2i1e40843_app6.png](#)]

References

1. Fouillet A, Fournet N, Caillère N. SurSaUD® Software: a tool to support the data management, the analysis and the dissemination of results from the french syndromic surveillance system. *Online J Public Health Informatics* 2013;5 [[FREE Full text](#)] [doi: [10.5210/ojphi.v5i1.4426](https://doi.org/10.5210/ojphi.v5i1.4426)]
2. Caserio SC, Henry V, Fouillet A, Bousquet V. Le Système de Surveillance Syndromique SurSaUDz. *Bulletin épidémiologique hebdomadaire* 2014;38-44 [[FREE Full text](#)]
3. Jossier L, Fouillet A, Caillère N, Brun-Ney D, Ille D, Brucker G, et al. Assessment of a syndromic surveillance system based on morbidity data: results from the Oscour network during a heat wave. *PLoS One* 2010 Aug 09;5:e11984 [[FREE Full text](#)] [doi: [10.1371/journal.pone.0011984](https://doi.org/10.1371/journal.pone.0011984)] [Medline: [20711252](https://pubmed.ncbi.nlm.nih.gov/20711252/)]
4. Paireau J, Pelat C, Caserio-Schönemann C, Pontais I, Le Strat Y, Lévy-Bruhl D, et al. Mapping influenza activity in emergency departments in France using Bayesian model-based geostatistics. *Influenza Other Respir Viruses* 2018 Nov 21;12:772-779 [[FREE Full text](#)] [doi: [10.1111/irv.12599](https://doi.org/10.1111/irv.12599)] [Medline: [30055089](https://pubmed.ncbi.nlm.nih.gov/30055089/)]

5. Hughes HE, Morbey R, Fouillet A, Caserio-Schönemann C, Dobney A, Hughes TC, et al. Retrospective observational study of emergency department syndromic surveillance data during air pollution episodes across London and Paris in 2014. *BMJ Open* 2018 Apr 19;8:e018732 [FREE Full text] [doi: [10.1136/bmjopen-2017-018732](https://doi.org/10.1136/bmjopen-2017-018732)] [Medline: [29674360](https://pubmed.ncbi.nlm.nih.gov/29674360/)]
6. Subiros M, Brottet E, Solet J, LeGuen A, Filleul L. Health monitoring during water scarcity in Mayotte, France, 2017. *BMC Public Health* 2019 Mar 12;19:288 [FREE Full text] [doi: [10.1186/s12889-019-6613-8](https://doi.org/10.1186/s12889-019-6613-8)] [Medline: [30866876](https://pubmed.ncbi.nlm.nih.gov/30866876/)]
7. GBD 2017 DALYsHALE Collaborators. Global, regional, and national disability-adjusted life-years (DALYs) for 359 diseases and injuries and healthy life expectancy (HALE) for 195 countries and territories, 1990-2017: a systematic analysis for the Global Burden of Disease Study 2017. *Lancet* 2018 Nov 10;392:1859-1922 [FREE Full text] [doi: [10.1016/S0140-6736\(18\)32335-3](https://doi.org/10.1016/S0140-6736(18)32335-3)] [Medline: [30415748](https://pubmed.ncbi.nlm.nih.gov/30415748/)]
8. Annual Report 2019-20. Department of Health & Family Welfare Ministry of Health & Family Welfare Government of India. 2019. URL: <https://main.mohfw.gov.in/sites/default/files/Annual%20Report%202019-2020%20English.pdf> [accessed 2020-03-03]
9. Global Burden of Disease (GBD). IHME. URL: <http://www.healthdata.org/gbd> [accessed 2020-03-01]
10. Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, et al. Attention is all you need. arXiv 2017. [doi: [10.48550/arXiv.1706.03762](https://doi.org/10.48550/arXiv.1706.03762)]
11. Devlin J, Chang M, Lee K, Toutanova K. BERT: pre-training of deep bidirectional transformers for language understanding. arXiv 2019 [FREE Full text]
12. Le H, Vial L, Frej J, Segonne V, Coavoux M, Lecouteux B, et al. FlauBERT: unsupervised language model pre-training for French. arXiv 2020 [FREE Full text]
13. Martin L, Muller B, Suárez P, Dupont Y, Romary L, de la Clergerie E, et al. CamemBERT: a tasty french language model. In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. 2019 Presented at: 58th Annual Meeting of the Association for Computational Linguistics; Jul, 2020; Online. [doi: [10.18653/v1/2020.acl-main.645](https://doi.org/10.18653/v1/2020.acl-main.645)]
14. Liu Y, Ott M, Goyal N, Du J, Joshi M, Chen D, et al. RoBERTa: a robustly optimized BERT pretraining approach. arXiv 2019 [FREE Full text]
15. Javier OS, Sagot B, Romary L. Asynchronous pipeline for processing huge corpora on medium to low resource infrastructures. In: Proceedings of the Workshop on Challenges in the Management of Large Corpora (CMCL-7) 2019. 2019 Presented at: Workshop on Challenges in the Management of Large Corpora (CMCL-7) 2019; Jul 22, 2019; Cardiff. [doi: [10.14618/IDS-PUB-9021](https://doi.org/10.14618/IDS-PUB-9021)]
16. Wenzek G, Lachaux M, Conneau A, Chaudhary V, Guzmán F, Joulin A, et al. CCNet: extracting high quality monolingual datasets from web crawl data. arXiv 2019 [FREE Full text]
17. Radford A, Wu J, Child R, Luan D, Amodei D, Sutskever I. Language models are unsupervised multitask learners. OpenAI blog. 2019. URL: https://d4mucfpxyww.cloudfront.net/better-language-models/language_models_are_unsupervised_multitask_learners.pdf [accessed 2022-12-15]
18. Louis A. BelGPT-2: a GPT-2 model pre-trained on French corpora. GitHub. 2021. URL: <https://github.com/antoiloui/belgpt2> [accessed 2022-12-15]
19. Suominen H, Kelly L, Goeriot L. Overview of the CLEF eHealth evaluation lab 2018. In: Experimental IR Meets Multilinguality, Multimodality, and Interaction. Cham: Springer; 2018.
20. Cossin S, Jouhet V, Mougin F, Diallo G, Thiessard F. IAM at CLEF eHealth 2018: concept annotation and coding in French death certificates. arXiv 2018 [FREE Full text]
21. Flicoteaux R. ECSTRA-APHP @ CLEF eHealth2018-task 1: ICD10 Code Extraction from Death Certificates. CEUR-WS. 2018. URL: https://ceur-ws.org/Vol-2125/paper_147.pdf [accessed 2022-12-15]
22. Amin-Nejad A, Ive J, Velupillai S. Exploring transformer text generation for medical dataset augmentation. In: Proceedings of the Twelfth Language Resources and Evaluation Conference. 2020 Presented at: Twelfth Language Resources and Evaluation Conference; May, 2020; Marseille, France URL: <https://aclanthology.org/2020.lrec-1.578/> [doi: [10.1007/978-1-4842-6150-7_7](https://doi.org/10.1007/978-1-4842-6150-7_7)]
23. Metzger M, Tvardik N, Gicquel Q, Bouvry C, Poulet E, Potinet-Pagliaroli V. Use of emergency department electronic medical records for automated epidemiological surveillance of suicide attempts: a French pilot study. *Int J Methods Psychiatr Res* 2017 Jun 15;26:e1522 [FREE Full text] [doi: [10.1002/mpr.1522](https://doi.org/10.1002/mpr.1522)] [Medline: [27634457](https://pubmed.ncbi.nlm.nih.gov/27634457/)]
24. Lopez-Garcia G, Jerez JM, Ribelles N, Alba E, Veredas FJ. Transformers for clinical coding in Spanish. *IEEE Access* 2021;9:72387-72397. [doi: [10.1109/access.2021.3080085](https://doi.org/10.1109/access.2021.3080085)]
25. Zhang Z, Liu J, Razavian N. BERT-XML: large scale automated ICD coding using BERT pretraining. In: Proceedings of the 3rd Clinical Natural Language Processing Workshop. 2020 Presented at: 3rd Clinical Natural Language Processing Workshop; Nov, 2020; Online URL: <http://arxiv.org/abs/2006.03685> [doi: [10.18653/v1/2020.clinicalnlp-1.3](https://doi.org/10.18653/v1/2020.clinicalnlp-1.3)]
26. Liu Z, He H, Yan S, Wang Y, Yang T, Li G. End-to-end models to imitate traditional Chinese medicine syndrome differentiation in lung cancer diagnosis: model development and validation. *JMIR Med Inform* 2020 Jun 16;8:e17821 [FREE Full text] [doi: [10.2196/17821](https://doi.org/10.2196/17821)] [Medline: [32543445](https://pubmed.ncbi.nlm.nih.gov/32543445/)]
27. Mohammadi R, Jain S, Namin AT, Scholem Heller M, Palacholla R, Kamarthi S, et al. Predicting unplanned readmissions following a hip or knee arthroplasty: retrospective observational study. *JMIR Med Inform* 2020 Nov 27;8:e19761 [FREE Full text] [doi: [10.2196/19761](https://doi.org/10.2196/19761)] [Medline: [33245283](https://pubmed.ncbi.nlm.nih.gov/33245283/)]

28. Landis JR, Koch GG. The measurement of observer agreement for categorical data. *Biometrics* 1977 Mar;33:159-174. [doi: [10.2307/2529310](https://doi.org/10.2307/2529310)]
29. Bird S. NLTK: the natural language toolkit. In: *Proceedings of the COLING/ACL on Interactive presentation sessions. 2006 Presented at: COLING-ACL '06: Proceedings of the COLING/ACL on Interactive presentation sessions; Jul 17 - 18, 2006; Sydney Australia* URL: <https://aclanthology.org/P06-4018.pdf> [doi: [10.3115/1225403.1225421](https://doi.org/10.3115/1225403.1225421)]
30. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, et al. *Scikit-learn: machine learning in Python*. arXiv 2018. [doi: [10.48550/arXiv.1201.0490](https://doi.org/10.48550/arXiv.1201.0490)]
31. Kudo T, Richardson J. SentencePiece: a simple and language independent subword tokenizer and detokenizer for Neural Text Processing. arXiv 2018 [FREE Full text] [doi: [10.18653/v1/d18-2012](https://doi.org/10.18653/v1/d18-2012)]
32. Shibata Y, Kida T, Fukamachi S, Takeda M, Shinohara A, Shinohara T, et al. Byte pair encoding: a text compression scheme that accelerates pattern matching. *ResearchGate*. 1999. URL: <https://tinyurl.com/56uv6tzb> [accessed 2022-12-15]
33. Komatsuzaki A. One epoch is all you need. arXiv 2019 [FREE Full text]
34. Grandini M, Bagli E, Visani G. Metrics for multi-class classification: an overview. arXiv 2020 [FREE Full text]
35. Özgür A, Özgür L, Güngör T. Text categorization with class-based and corpus-based keyword selection. In: *Computer and Information Sciences - ISICIS 2005*. ISICIS 2005. Berlin, Heidelberg: Springer; 2005.
36. Wu S, Roberts K, Datta S, Du J, Ji Z, Si Y, et al. Deep learning in clinical natural language processing: a methodical review. *J Am Med Inform Assoc* 2020 Mar 01;27:457-470 [FREE Full text] [doi: [10.1093/jamia/ocz200](https://doi.org/10.1093/jamia/ocz200)] [Medline: [31794016](https://pubmed.ncbi.nlm.nih.gov/31794016/)]
37. Mascio A, Kraljevic Z, Bean D, Dobson R, Stewart R, Bendayan R, et al. Comparative analysis of text classification approaches in electronic health records. arXiv 2020 [FREE Full text] [doi: [10.18653/v1/2020.bionlp-1.9](https://doi.org/10.18653/v1/2020.bionlp-1.9)]
38. Yao L, Jin Z, Mao C, Zhang Y, Luo Y. Traditional Chinese medicine clinical records classification with BERT and domain specific corpora. *J Am Med Inform Assoc* 2019 Dec 01;26:1632-1636 [FREE Full text] [doi: [10.1093/jamia/ocz164](https://doi.org/10.1093/jamia/ocz164)] [Medline: [31550356](https://pubmed.ncbi.nlm.nih.gov/31550356/)]
39. Lee J, Hsiang J. Patent claim generation by fine-tuning OpenAI GPT-2. *World Patent Inform* 2020 Sep;62:101983. [doi: [10.1016/j.wpi.2020.101983](https://doi.org/10.1016/j.wpi.2020.101983)]
40. Huang K, Altsaar J, Ranganath R. ClinicalBERT: modeling clinical notes and predicting hospital readmission. arXiv 2019 Apr 10 [FREE Full text]
41. Xu B, Gil-Jardiné C, Thiessard F, Tellier E, Avalos M, Lagarde E. Pre-training a neural language model improves the sample efficiency of an emergency room classification model. arXiv 2021 Apr 7. [doi: [10.32473/flairs.v34i1.128480](https://doi.org/10.32473/flairs.v34i1.128480)]
42. Adams G, Ketenci M, Bhave S, Perotte A, Elhadad N. Zero-shot clinical acronym expansion via latent meaning cells. arXiv 2020 Nov [FREE Full text]
43. Egan N, Bohannon J. Primer AI's systems for acronym identification and disambiguation. arXiv 2021 Jan [FREE Full text]
44. Martin L, Muller B, Suárez PJ, Dupont Y, Romary L, de la Clergerie EV, et al. Les modèles de langue contextuels Camembert pour le français : impact de la taille et de l'hétérogénéité des données d'entraînement (C AMEM BERT Contextual Language Models for French: Impact of Training Data Size and Heterogeneity). In: *Actes de la 6e conférence conjointe Journées d'Études sur la Parole (JEP, 33e édition), Traitement Automatique des Langues Naturelles (TALN, 27e édition), Rencontre des Étudiants Chercheurs en Informatique pour le Traitement Automatique des Langues (RÉCITAL, 22e édition)*. Volume 2 : Traitement Automatique des Langues Naturelles. Nancy, France: ATALA et AFCEP; 2020.
45. WHO Guide to Identifying the Economic Consequences of Disease and Injury. Geneva: Department of Health Systems Financing Health Systems and Services World Health Organization; 2009.
46. Krug EG, Sharma GK, Lozano R. The global burden of injuries. *Am J Public Health* 2000 Apr 01;90:523-526. [doi: [10.2105/ajph.90.4.523](https://doi.org/10.2105/ajph.90.4.523)] [Medline: [10754963](https://pubmed.ncbi.nlm.nih.gov/10754963/)]
47. Injury Surveillance Guidelines. Geneva: World Health Organization; Mar 16, 2001.
48. Peden M. *World Report on Road Traffic Injury Prevention Summary*. Darby, PA, U.S.A: DIANE Publishing Company; 2008.

Abbreviations

BERT: Bidirectional Encoder Representations Transformer

CNN: convolutional neural network

ED: emergency department

EHR: electronic health record

FEDORU: Fédération des Observatoires Régionaux des Urgences

GPT-2: Generative Pretrained Transformer-2

ICD: International Classification of Diseases

MVA: motor vehicle accident

SurSaUD: Surveillance Sanitaire des Urgences et des Décès

SVM: support vector machine

TARPON: Traitement Automatique des Résumés de Passage aux urgences dans le but de créer un Observatoire National

TF-IDF: term frequency–inverse document frequency

Edited by K El Emam, B Malin; submitted 07.07.22; peer-reviewed by S Matos, Z Li; comments to author 02.10.22; revised version received 14.10.22; accepted 29.10.22; published 12.01.23.

Please cite as:

*Chenais G, Gil-Jardiné C, Touchais H, Avalos Fernandez M, Contrand B, Tellier E, Combes X, Bourdois L, Revel P, Lagarde E
Deep Learning Transformer Models for Building a Comprehensive and Real-time Trauma Observatory: Development and Validation Study*

JMIR AI 2023;2:e40843

URL: <https://ai.jmir.org/2023/1/e40843>

doi: [10.2196/40843](https://doi.org/10.2196/40843)

PMID: [38875539](https://pubmed.ncbi.nlm.nih.gov/38875539/)

©Gabrielle Chenais, Cédric Gil-Jardiné, Hélène Touchais, Marta Avalos Fernandez, Benjamin Contrand, Eric Tellier, Xavier Combes, Loick Bourdois, Philippe Revel, Emmanuel Lagarde. Originally published in JMIR AI (<https://ai.jmir.org>), 12.01.2023. This is an open-access article distributed under the terms of the Creative Commons Attribution License (<https://creativecommons.org/licenses/by/4.0/>), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR AI, is properly cited. The complete bibliographic information, a link to the original publication on <https://www.ai.jmir.org/>, as well as this copyright and license information must be included.

Original Paper

Few-Shot Learning for Clinical Natural Language Processing Using Siamese Neural Networks: Algorithm Development and Validation Study

David Oniani¹; Premkumar Chandrasekar¹; Sonish Sivarajkumar², BSc; Yanshan Wang^{1,2,3,4}, PhD

¹Department of Health Information Management, University of Pittsburgh, Pittsburgh, PA, United States

²Intelligent Systems Program, University of Pittsburgh, Pittsburgh, PA, United States

³Department of Biomedical Informatics, University of Pittsburgh, Pittsburgh, PA, United States

⁴Clinical and Translational Science Institute, University of Pittsburgh, Pittsburgh, PA, United States

Corresponding Author:

Yanshan Wang, PhD

Department of Health Information Management

University of Pittsburgh

6026 Forbes Tower

Pittsburgh, PA, 15260

United States

Phone: 1 4123832712

Email: yanshan.wang@pitt.edu

Abstract

Background: Natural language processing (NLP) has become an emerging technology in health care that leverages a large amount of free-text data in electronic health records to improve patient care, support clinical decisions, and facilitate clinical and translational science research. Recently, deep learning has achieved state-of-the-art performance in many clinical NLP tasks. However, training deep learning models often requires large, annotated data sets, which are normally not publicly available and can be time-consuming to build in clinical domains. Working with smaller annotated data sets is typical in clinical NLP; therefore, ensuring that deep learning models perform well is crucial for real-world clinical NLP applications. A widely adopted approach is fine-tuning existing pretrained language models, but these attempts fall short when the training data set contains only a few annotated samples. Few-shot learning (FSL) has recently been investigated to tackle this problem. Siamese neural network (SNN) has been widely used as an FSL approach in computer vision but has not been studied well in NLP. Furthermore, the literature on its applications in clinical domains is scarce.

Objective: The aim of our study is to propose and evaluate SNN-based approaches for few-shot clinical NLP tasks.

Methods: We propose 2 SNN-based FSL approaches, including pretrained SNN and SNN with second-order embeddings. We evaluate the proposed approaches on the clinical sentence classification task. We experiment with 3 few-shot settings, including 4-shot, 8-shot, and 16-shot learning. The clinical NLP task is benchmarked using the following 4 pretrained language models: bidirectional encoder representations from transformers (BERT), BERT for biomedical text mining (BioBERT), BioBERT trained on clinical notes (BioClinicalBERT), and generative pretrained transformer 2 (GPT-2). We also present a performance comparison between SNN-based approaches and the prompt-based GPT-2 approach.

Results: In 4-shot sentence classification tasks, GPT-2 had the highest precision (0.63), but its recall (0.38) and *F* score (0.42) were lower than those of BioBERT-based pretrained SNN (0.45 and 0.46, respectively). In both 8-shot and 16-shot settings, SNN-based approaches outperformed GPT-2 in all 3 metrics of precision, recall, and *F* score.

Conclusions: The experimental results verified the effectiveness of the proposed SNN approaches for few-shot clinical NLP tasks.

(JMIR AI 2023;2:e44293) doi:[10.2196/44293](https://doi.org/10.2196/44293)

KEYWORDS

few-shot learning; FSL; Siamese neural network; SNN; natural language processing; NLP; neural networks

Introduction

Background

Deep neural networks (DNNs), due to their performance [1], currently dominate both computer vision and natural language processing (NLP) literature. However, fully using the capabilities of DNNs requires large training data sets. To tackle this problem, researchers have tried to reduce the complexity of the DNN models to obtain comparable performance when the training data set is small [2]. The few-shot learning (FSL) paradigm is an alternative attempt that aims to improve model performance under data constraints. The goal of FSL is to efficiently learn from a small number of *shots* (ie, data samples or instances). The number of samples usually ranges from 1 to 100 per class [3,4]. There is a growing interest in the artificial intelligence (AI) research community in FSL, and several different strategies have been developed for FSL, including Bowtie Networks [5], Induction Networks [6], and Prototypical Networks [7].

A Siamese neural network (SNN), sometimes called a twin neural network, is an artificial neural network that uses 2 parallel, weight-sharing machine learning models to compute comparable embeddings. The SNN architecture has shown promising results as an FSL approach in computer vision for similarity detection [8] and duplicate identification [9]. Yet, its usage in NLP has been understudied, and, to the best of our knowledge, there have not been any studies investigating SNNs for clinical NLP.

In SNNs, neural networks are trained to compute embeddings. In NLP, deep learning has achieved state-of-the-art performance since it could generate comprehensive embeddings to encode semantic and syntactic information. The primary use of deep learning in NLP is to represent the language in a vectorized form (ie, embeddings) so that the representation can be used for different NLP tasks, such as natural language generation, text classification, and semantic textual similarity. Thus, having a robust embedding-generation mechanism is crucial for most NLP tasks. Since the context of words, sentences, and more generally, text is important to learn meaningful embeddings, context-aware embedding-generation models, such as bidirectional encoder representations from transformers (BERT) [10], often show promising results. Furthermore, depending on the domain, the context also varies. For this purpose, researchers and engineers have built domain-specific, specialized models for use in downstream tasks. Examples of such models include BERT for biomedical text mining (BioBERT) [11] trained from biomedical literature texts and Bio + clinical BERT (BioClinicalBERT) trained from clinical texts [12]. However, leveraging contextual embeddings for FSL has rarely been studied in clinical NLP.

FSL is critical for clinical NLP as annotating a large training data set is costly and usually requires involving domain experts. On the other hand, it is common to have a few clinical text samples annotated by physicians. One example could be clinical notes with annotations of a rare disease, with the number of samples limited due to the nature of the disease. Despite such challenges, the importance of using AI in clinical applications

cannot be understated. AI could assist physicians in their decision-making, facilitate clinical and translational research, and significantly reduce the need for manual work. This study proposes an FSL approach based on SNNs to tackle clinical NLP tasks with only a few annotated training samples. Two SNN-based FSL approaches are proposed: pretrained SNN (PT-SNN) and SNN with second-order embeddings (SOE-SNN). Both approaches used the 3 different transformer models of BERT, BioBERT, and BioClinicalBERT. We evaluated the proposed strategies on the clinical sentence classification task. Clinical text classification refers to the classification of clinical sentences based on predefined classes. We show that SNN-based methods outperform the baseline, generative pretrained transformer 2 (GPT-2) model in few-shot settings for the task. Finally, we discuss the limitations and future work.

Related Work

There have been studies evaluating the usability of SNNs for image classification. Li et al [13] used SNNs for the classification of high-dimensional radiomic features extracted from MRI images. Hunt et al [14] applied SNNs for the classification of electrograms. Zhao et al [15] have used SNNs for hyperspectral image classification.

In sentence classification, Reimers and Gurevych [16] used SNNs to derive semantically meaningful sentence embeddings that can be compared using cosine similarity. It is important to note that the package we used in our experiments to generate embeddings was based on this paper [16]. However, the primary goal of our experiments was not generating sentence embeddings, but rather designing techniques for using such embeddings in few-shot clinical sentence classification tasks.

In the context of FSL, SNNs have been used by Torres et al [17] for one-shot, convolutional neural networks-based classification to optimize the discovery of novel compounds based on a reduced set of candidate drugs. Droghini et al [18] employed SNNs for few-shot human fall detection purposes using images. However, none of these studies used SNN-based FSL for NLP.

In few-shot text classification, Wei et al [19] used data augmentation to improve the performance of triplet networks. Liu et al [20] proposed distribution estimation to augment the labeled samples by sampling from the estimated distribution. Wang et al [21] represented each task using gradient information from a base model and trained an adaptation network that modulates a text classifier conditioned on the task representation.

There is only a recent study by Müller et al [22] that explored SNNs for FSL in NLP and demonstrated the high performance of pretrained SNNs that embed texts and labels. To the best of our knowledge, none of the studies referenced above are using SNNs to perform FSL in the clinical NLP domain.

Methods

Ethical Considerations

As the study is using a publicly available data set that is accessible under the data use agreement, there is no requirement for an institutional review board.

Data Set Derived from the Medical Information Mart for Intensive Care

The sentences were obtained from the Medical Information Mart for Intensive Care (MIMIC-III) database [23]. We used the same data set as in the HealthPrompt paper by Sivarajkumar and Wang [24], but with classes suitable for 4-shot, 8-shot, and 16-shot FSL experiments. In total, the data set had 444 samples

and 4 classes. Table 1 shows the distribution of classes in the data set.

Since we had 444 samples in total and performed 4-, 8-, and 16-shot experiments, the train size varied and was 16, 32, and 64 samples with the test sizes of 428, 412, and 380 samples, respectively.

Table 1. Few-shot sentence classification data set (N=444).

Label	Sample, n (%)
ADVANCED.LUNG.DISEASE	245 (55.2)
ADVANCED.HEART.DISEASE	117 (26.4)
CHRONIC.PAIN.FIBROMYALGIA	48 (10.8)
ADVANCED.CANCER	34 (7.7)

Sentence-Level Embeddings

For generating contextual, sentence-level embeddings, we used the sentence-transformers package [25]. The package provides intuitive and easy-to-use methods for computing dense vector representations of sentences, paragraphs, and images. The models are based on transformers such as BERT, RoBERTa [26], and so on, and achieve state-of-the-art performance in various tasks. The generated embeddings are such that similar texts are close in the latent space and can efficiently be found using cosine similarity. Thus, for sentences a and b with the corresponding embeddings A and B , we can compute the cosine similarity as follows:

$$\text{cosine similarity}(A, B) = (A \cdot B) / (\|A\|_2 \|B\|_2) \quad (1)$$

Model Architecture

The SNN's architecture leverages 2 parallel weight-sharing machine learning models (Figure 1). In the forward pass, 2 samples are passed into the models and mapped down to the latent space. The embeddings in the latent space are then compared using a similarity function, as shown in Equation 2. The similarity function is a hyperparameter that can vary based on the task and could range from Euclidean distance to Manhattan distance or cosine similarity. Depending on the

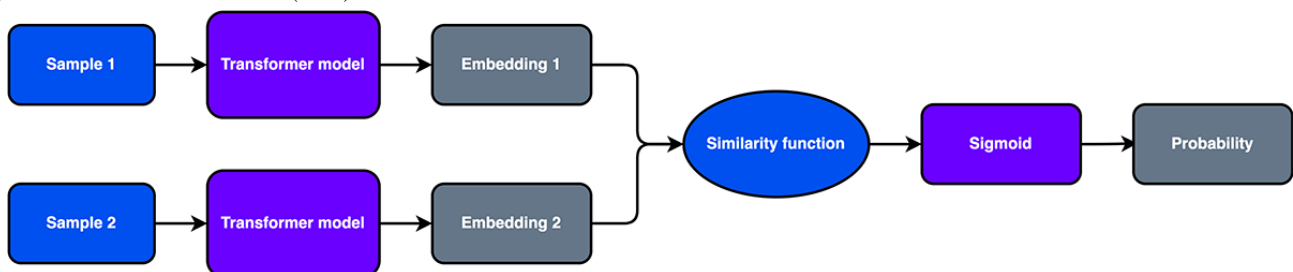
similarity function, the similarity value can then be mapped onto the (0, 1) interval by applying the Sigmoid function. Finally, a high similarity value means that the input samples likely belong to the same category and vice versa.

$$\text{out} = \sigma(\text{distance}(\text{emb}_1, \text{emb}_2)) \quad (2)$$

During training, SNN conducts representation learning [27] and attempts to have the best approximation for the input embeddings. The representation is learned by penalizing the loss if the model yields a high similarity value for inputs from different classes or if the model yields a low similarity value for inputs from the same class.

The SNN architecture naturally allows for data augmentation. For instance, in the case of 8-shot learning, the traditional training approach would involve passing 8 samples directly into the model. This approach is very limiting with such a small number of samples. SNN takes a different route and instead considers unique comparisons within the training set. With the training set consisting of 8 samples, there are $8 * 7 / 2 = 28$ unique comparisons. Thus, instead of 8 training samples, we get 28, which is 3.5 times more. In the case of 16 samples, the improvement is even more significant as the number of unique comparisons is 120, and there is a 7.5-fold data augmentation.

Figure 1. Siamese neural network (SNN) architecture.



More generally, under N -way- K -shot classification settings, for the data set D_{train} with N class labels and K labeled samples for each class, the following holds after SNN-style augmentation:

$$D_{\text{train SNN}} = \{(x_i, x_j) \mid x_i, x_j \in D_{\text{train}}, i < j\} \quad (3)$$

$$\text{size}(D_{\text{train SNN}}) = ((NK)^2 - NK) / 2 \quad (4)$$

Pretrained SNN

In the first approach, we leverage the pretrained language models (PLMs) to generate embeddings for the SNN, called pretrained SNN (PT-SNN). We used 3 PLMs in this approach, namely BERT, BioBERT, and BioClinicalBERT, to generate embeddings for the input training samples.

In the following, we illustrate how to use the PT-SNN for classification. Suppose we want to perform binary classification. We are given 2 classes C_1 and C_2 , a training set D_{train} , and a testing set D_{test} . We first compute embeddings for all samples in both D_{train} and D_{test} . For every testing sample, using the generated embeddings, we compute the similarity with respect to every training sample and compute the mean similarity values for classes. For instance, mean similarity value for some samples $x \in D_{test}$ with respect to C_1 and C_2 might be 0.2 and 0.6, respectively. Since 0.6 is greater than 0.2, we classify sample

x as being in class C_2 . It should be noted that the algorithm is similar to the k-nearest neighbors [28] classification algorithm.

Note that our classification approach is such that using Sigmoid is not necessary. In the case of SOE-SNN, it is required during training, but not during testing (See Algorithm 1 in [Textbox 1](#)).

Algorithm 1 presents the pseudocode for the classification algorithm and the evaluation approach. Here, *EvalIters* refers to the number of averaging iterations for addressing the instability issues. In our case, *EvalIters* is 3.

Textbox 1. Algorithm 1—our proposed algorithm for Siamese neural network–style classification and evaluation for few-shot learning.

```

Require:  $D_{train}$ : Train data set
Require:  $E_{test}$ : Test data set embeddings
Require:  $L_{test}$ : Test data set labels
Require: EvalIters: Number of evaluation iterations
Require: RandSubset: A function that randomly subsets a data set with the given seed
Require: L1Normalize: L1-normalizes the input tensor
Require: L2Normalize: L2-normalizes the input tensor
Require: Arange: Constructs a tensor of numbers from the given start and end (exclusive) with the step size of one
Require: Argmax: Finds the index of the maximum value along the given dimension
Require: ComputeMetrics: Computes evaluation metrics: precision, recall, and  $F$  score
Require: MatMul: Performs a matrix multiplication of the given tensors
Require: Max: Finds the maximum value of all elements in the input tensor
Require: Mean: Calculates the mean of a vector along the given dimension
Require: NumElements: Finds the number of elements in the input tensor
Require: Transpose: Transposes the input tensor
Require: Zeros: Creates the tensor of zeros with the given dimensions
1:  $Metrics \leftarrow Zeros(EvalIters, 3)$ ;
2: for  $Idx \leftarrow 0$  to EvalIters do
3:  $E_{train}, L_{train} \leftarrow RandSubset(D_{train}, Seed = Idx)$ ;
4:  $L2_{train} \leftarrow L2Normalize(E_{test})$ ;
5:  $L2_{test} \leftarrow L2Normalize(E_{train})$ ;
6:  $SimilarityTable \leftarrow MatMul(L2_{test}, Transpose(L2_{train}))$ ;
7:  $LabelTable \leftarrow Zeros(Max(L_{train}) + 1, NumElements(L_{train}))$ ;
8:  $LabelTable [L_{train}, Arange(NumElements(L_{train}))] \leftarrow 1$ ;
9:  $LabelTable \leftarrow L1Normalize(LabelTable)$ ;
10:  $Out \leftarrow Argmax(MatMul(SimilarityTable, Transpose(LabelTable)), Dim = 1)$ ;
11:  $Metrics[Idx] \leftarrow ComputeMetrics(L_{test}, Out)$ ;
12: end for
13:  $Precision, Recall, Fscore \leftarrow Mean(Metrics, Dim = 0)$ .

```

We use the vectorized implementations of cosine similarity, group by, and aggregate operations described in [Multimedia Appendix 1](#).

Such a strategy for classification can be slow in cases where the training set is large. However, the proposed approach is

feasible in the FSL settings, where the number of annotated samples is limited. Thus, we do not expect significant performance drawbacks when the number of samples is not large. Furthermore, the proposed PT-SNN approach can be high-performing under FSL settings.

We have also released a codebase implementing the proposed algorithms and models [29].

SNN With Second-Order Embeddings

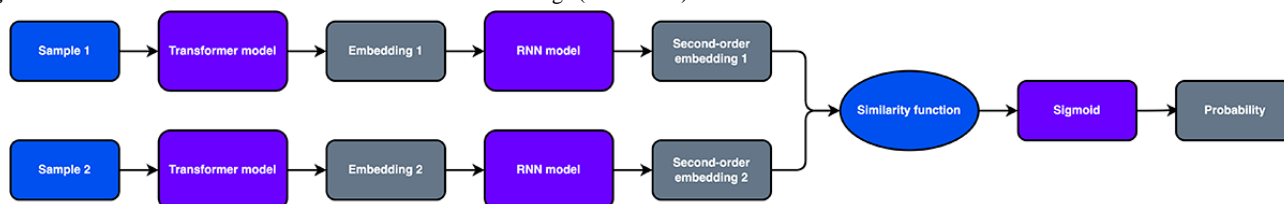
The second proposed approach is SOE-SNN where we apply an additional recurrent neural network (RNN) layer, such as long-short term memory or gated recurrent unit to the generated embeddings and then train the SNN model in the fashion described in the model architecture section (Figure 2). In our experiments, we used bidirectional long-short term memory for producing second-order embeddings.

Specifically, we first obtain the embeddings for all training samples from the PLMs. Half of the samples are used for training the RNN, and the other half is used for the classification

algorithm described in Algorithm 1. For the RNN half, all possible unique pairs of the samples are generated and labeled 1 if the samples in the pair are of the same class or 0 if they come from different classes. Binary cross entropy [30] and AdamW [31] are used as the loss function and the optimizer, respectively. The loss function and the optimizer were used for training the RNN. Similar to the PT-SNN, the transformer model that generates embeddings was not updated, and as such, one could think of this as a frozen component of the training pipeline.

Model evaluation is done in the same manner as in PT-SNNs, where we compute mean similarity scores and average out the metrics over 3 evaluation iterations to handle the potential instability issues.

Figure 2. Siamese neural network with second-order embeddings (SOE-SNN) architecture. RNN: recurrent neural network.



FSL Model Evaluation

Systematically evaluating FSL model performance can be tricky since fine-tuning or making predictions on small data sets could potentially suffer from instability [32]. To address this issue, we propose the averaging strategy for model evaluation. For every few-shot experiment (eg, 4-shot, 8-shot, and 16-shot experiments), we use randomized sampling to sample 4, 8, or 16 samples per class and create a training data set. We perform this M times, and therefore, for every experiment, M randomly generated training sets are evaluated on the test set. Finally, the metrics are averaged out and reported as the final scores.

$$\text{Metric} = (\sum_{i=1}^M \text{Metric}_i) / M \quad (5)$$

Such an approach gives a more robust view of the model's performance in possibly unstable scenarios. Therefore, we choose $M=3$ and employ this strategy in all reported metrics. As for metrics, we choose precision, recall, and F score.

Baseline Model

Despite the availability of newer GPT models such as ChatGPT and GPT-4, they cannot be used on the MIMIC data set as per the terms of the data use agreement. Therefore, we used the open source GPT-2. We used the GPT-2 [33] with 355 million parameters as the baseline model. We obtained 4, 8, and 16 samples per class to generate predictions. To achieve this, we used the transformers package [34]. Note that no fine-tuning was done in this case, and instead, the existing GPT-2 model was used directly for generating responses.

We used a prefix prompt with all possible classes appended to the sentence for classification, followed by the incomplete sentence that would have to be completed by GPT-2. The proposed prompt is similar to the cloze prompt that showed the best performance in Sivarajkumar and Wang [24]. We modified the prompt by adding additional information at the end of the text (all 4 labels) and moved the mask at the end, effectively

turning it into a prefix prompt. Thus, we used the following prompt:

```
{text}. options are advanced cancer, advanced heart disease, advanced lung disease, chronic pain fibromyalgia. type of disease {mask}
```

where {text} is the input text and {mask} is the placeholder for GPT-2 to fill in with the generated text. Appending the list of labels to the end of the input text was done to help the GPT-2 model by showing all available options. We used the maximum context size of 1024—the most GPT-2 can handle. If the total number of tokens exceeded 1024, the sentence was trimmed from the end to keep the prompt intact.

Finally, the generated responses were analyzed and evaluated by the annotator. The annotator labeled every GPT-2 response with the semantically closest class (1 of 4 options). Note that the annotator evaluated the responses only once. Thus, for GPT-2, the number of averaging iterations is 1 (ie, $M=1$).

Results

We present the results of 4-shot, 8-shot, and 16-shot experiments for few-shot sentence classification task. We used models based on BERT, BioBERT, BioClinicalBERT. The results are shown in Table 2.

In the 4-shot sentence classification task, the baseline, GPT-2 model had the highest precision (0.63). BioClinicalBERT-based SOE-SNN came next with a precision score of 0.57. PT-SNN had the highest recall and F score values of 0.45 and 0.46, respectively. BioClinicalBERT-based PT-SNN was the second with recall and F score of 0.42 and 0.43, respectively. Thus, in 4-shot settings, GPT-2 had a higher precision, but its recall and F score were lower than those of SNN-based approaches.

In 8-shot experiments, BioClinicalBERT-based PT-SNN outperformed all other approaches in precision, with a value of

0.64. BioBERT-based SOE-SNN had both the highest recall and the highest F score of 0.50 and 0.53, respectively. GPT-2 did not have the highest score in any of the metrics. Hence, for 8-shot learning, SNN-based approaches outperformed GPT-2.

As for 16-shot learning, BioClinicalBERT-based SOE-SNN had the highest precision value of 0.70. BioBERT-based

PT-SNN had the highest recall (0.55), and BioClinicalBERT-based PT-SNN had the highest F score (0.58). GPT-2 did not have the highest score in any of the metrics, with most models having higher precision, recall, and F score. Overall, SNN-based approaches outperformed the baseline GPT-2 model.

Table 2. Few-shot sentence classification.

Approach	Model	Shots	Precision	Recall	F score
GPT-2 ^a	GPT-2	4	0.63	0.38	0.42
PT-SNN ^b	BERT ^c	4	0.49	0.37	0.37
PT-SNN	BioBERT ^d	4	0.53	0.45	0.46
PT-SNN	BioClinicalBERT ^e	4	0.50	0.42	0.43
SOE-SNN ^f	BERT	4	0.49	0.26	0.30
SOE-SNN	BioBERT	4	0.52	0.19	0.17
SOE-SNN	BioClinicalBERT	4	0.57	0.27	0.24
GPT-2	GPT-2	8	0.63	0.38	0.42
PT-SNN	BERT	8	0.62	0.45	0.47
PT-SNN	BioBERT	8	0.61	0.48	0.50
PT-SNN	BioClinicalBERT	8	0.64	0.44	0.49
SOE-SNN	BERT	8	0.55	0.43	0.46
SOE-SNN	BioBERT	8	0.61	0.50	0.53
SOE-SNN	BioClinicalBERT	8	0.58	0.32	0.32
GPT-2	GPT-2	16	0.65	0.38	0.42
PT-SNN	BERT	16	0.64	0.51	0.52
PT-SNN	BioBERT	16	0.65	0.55	0.56
PT-SNN	BioClinicalBERT	16	0.69	0.54	0.58
SOE-SNN	BERT	16	0.59	0.44	0.48
SOE-SNN	BioBERT	16	0.43	0.38	0.36
SOE-SNN	BioClinicalBERT	16	0.70	0.39	0.38

^aGPT-2: generative pretrained transformer 2.

^bPT-SNN: pretrained Siamese neural network.

^cBERT: bidirectional encoder representations from transformers.

^dBioBERT: bidirectional encoder representations from transformers for biomedical text mining.

^eBioClinicalBERT: Bio + clinical bidirectional encoder representations from transformers.

^fSOE-SNN: Siamese neural network with second-order embeddings.

Discussion

Limitations and Future Work

There are several limitations of the work that can be addressed by further exploring FSL and SNNs. First, we did not compare the results to traditional baseline models such as support vector machine, logistic regression, multinomial logistic regression, random forest, and so on. Second, other data sets could also be used for evaluating the performance of SNNs in text classification. Third, since we can perform sentence-level classification, another interesting research direction could be document classification, where a document can be modeled as

a collection of sentences. Fourth, in the SOE-SNN, since we only experiment with 1 splitting strategy (half for fine-tuning embeddings and half for classification and evaluation), other RNN training versus classification ratios can also be noteworthy. Additionally, it is important to note that data sets for FSL, especially clinical FSL, are difficult to find. Ge et al [35] have emphasized that “(68%) studies reconstructed existing datasets to create few-shot scenarios synthetically.” Thus, building a brand-new FSL data set and then evaluating the performance of the proposed methods could also be an interesting future research direction.

Conclusion

We conducted few-shot learning experiments evaluating the performance of SNN models on the clinical sentence classification task. The SNN models were based on transformer models—BERT, BioBERT, and BioClinicalBERT. Since

performance evaluation on small data sets may suffer from instability, a special evaluation strategy was used. We conclude that, overall, SNN-based models outperformed the baseline GPT-2 model for sentence classification tasks. The limitations of the work have also been discussed alongside potential future directions of research.

Acknowledgments

The authors would like to acknowledge support from the University of Pittsburgh Momentum Funds, Clinical and Translational Science Institute Exploring Existing Data Resources Pilot Awards, the School of Health and Rehabilitation Sciences Dean's Research and Development Award, and the National Institutes of Health through Grant UL1TR001857.

Conflicts of Interest

None declared.

Multimedia Appendix 1

Vectorized cosine similarity, group by, and aggregate.

[[DOC File , 36 KB - ai_v2i1e44293_app1.doc](#)]

References

1. Sejnowski TJ. The unreasonable effectiveness of deep learning in artificial intelligence. *Proc Natl Acad Sci U S A* 2020 Dec 01;117(48):30033-30038 [[FREE Full text](#)] [doi: [10.1073/pnas.1907373117](https://doi.org/10.1073/pnas.1907373117)] [Medline: [31992643](https://pubmed.ncbi.nlm.nih.gov/31992643/)]
2. Brigato L, Iocchi L. A close look at deep learning with small data. 2021 Presented at: 25th International Conference on Pattern Recognition (ICPR); January 10-15, 2021; Milan, Italy. [doi: [10.1109/icpr48806.2021.9412492](https://doi.org/10.1109/icpr48806.2021.9412492)]
3. Mo Y, Xiaoxiao G, Jinfeng Y, Shiyu C, Saloni P, Yu C, et al. Diverse few-shot text classification with multiple metrics. 2018 Presented at: Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers); June 1-6, 2018; New Orleans, Louisiana, USA. [doi: [10.18653/v1/n18-1109](https://doi.org/10.18653/v1/n18-1109)]
4. Emmanouil M, Sepideh M, Alessandro B, Selene B, Robert JS. Give it a shot: Few-shot learning to normalize ADR mentions in social media posts. 2019 Presented at: Fourth Social Media Mining for Health Applications (#SMM4H) Workshop & Shared Task; August 2, 2019; Florence, Italy. [doi: [10.18653/v1/w19-3219](https://doi.org/10.18653/v1/w19-3219)]
5. Zhipeng B, Yu-Xiong W, Martial H. Bowtie networks: Generative modeling for joint few-shot recognition and novel-view synthesis. 2021 Presented at: International Conference on Learning Representations; May 3-7, 2021; Virtual. [doi: [10.48550/arXiv.2008.06981](https://doi.org/10.48550/arXiv.2008.06981)]
6. Ruiying G, Binhua L, Yongbin L, Xiaodan Z, Ping J, Jian S. Induction networks for few-shot text classification. 2019 Presented at: The 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP); November 3-7, 2019; Hong Kong, China. [doi: [10.18653/v1/d19-1403](https://doi.org/10.18653/v1/d19-1403)]
7. Jake S, Swersky K, Zemel RS. Prototypical networks for few-shot learning. *arXiv*. 2017:1-13. [doi: [10.48550/arXiv.1703.05175](https://doi.org/10.48550/arXiv.1703.05175)]
8. Wu Y, Wang W. Code similarity detection based on siamese network. 2021 Presented at: IEEE International Conference on Information Communication and Software Engineering (ICICSE); March 19-21, 2021; Chengdu, China. [doi: [10.1109/icicse52190.2021.9404110](https://doi.org/10.1109/icicse52190.2021.9404110)]
9. Fisichella M. Siamese coding network and pair similarity prediction for near-duplicate image detection. *Int J Multimed Info Retr* 2022 Apr 12;11(2):159-170. [doi: [10.1007/s13735-022-00233-w](https://doi.org/10.1007/s13735-022-00233-w)]
10. Jacob D, Ming-Wei C, Kenton L, Kristina T. BERT: Pre-training of deep bidirectional transformers for language understanding. 2018 Presented at: Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers); June 1-6, 2018; New Orleans, Louisiana, USA. [doi: [10.18653/v1/n18-2](https://doi.org/10.18653/v1/n18-2)]
11. Lee J, Yoon W, Kim S, Kim D, Kim S, So CH, et al. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics* 2020 Feb 15;36(4):1234-1240 [[FREE Full text](#)] [doi: [10.1093/bioinformatics/btz682](https://doi.org/10.1093/bioinformatics/btz682)] [Medline: [31501885](https://pubmed.ncbi.nlm.nih.gov/31501885/)]
12. Emily A, John M, William B, Wei-Hung W, Di J, Tristan N, et al. Publicly available clinical BERT embeddings. 2019 Presented at: The 2nd Clinical Natural Language Processing Workshop; June 7, 2019; Minneapolis, Minnesota, USA. [doi: [10.18653/v1/w19-1909](https://doi.org/10.18653/v1/w19-1909)]

13. Mahajan A, Dormer J, Li Q, Chen D, Zhang Z, Fei B. Siamese neural networks for the classification of high-dimensional radiomic features. *Proc SPIE Int Soc Opt Eng* 2020 Feb;11314:2020 [FREE Full text] [doi: [10.1117/12.2549389](https://doi.org/10.1117/12.2549389)] [Medline: [32528215](https://pubmed.ncbi.nlm.nih.gov/32528215/)]
14. Hunt B, Kwan E, Dossall D, MacLeod RS, Ranjan R. Siamese neural networks for small dataset classification of electrograms. 2021 Presented at: *Computing in Cardiology (CinC)*; September 13-15, 2021; Brno, Czech Republic. [doi: [10.23919/cinc53138.2021.9662707](https://doi.org/10.23919/cinc53138.2021.9662707)]
15. Zhao S, Li W, Du Q, Ran Q. Hyperspectral classification based on siamese neural network using spectral-spatial feature. 2018 Presented at: *2018 IEEE International Geoscience and Remote Sensing Symposium*; July 22-27, 2018; Valencia, Spain. [doi: [10.1109/igarss.2018.8519286](https://doi.org/10.1109/igarss.2018.8519286)]
16. Reimers N, Gurevych I. Sentence-bert: Sentence embeddings using siamese bert-networks. 2019 Presented at: *Proceedings of the Conference on Empirical Methods in Natural Language Processing*; November 3-7, 2019; Hong Kong, China. [doi: [10.18653/v1/d19-1410](https://doi.org/10.18653/v1/d19-1410)]
17. Torres L, Monteiro N, Oliveira J, Arrais J, Ribeiro B. Exploring a siamese neural network architecture for one-shot drug discovery. 2020 Presented at: *2020 IEEE 20th International Conference on Bioinformatics and Bioengineering (BIBE)*, pages 168-175, 2020; October 26-28, 2020; Cincinnati, OH, USA. [doi: [10.1109/bibe50027.2020.00035](https://doi.org/10.1109/bibe50027.2020.00035)]
18. Droghini D, Vesperini F, Principi E, Squartini S, Piazza F. Few-shot siamese neural networks employing audio features for human-fall detection. 2018 Presented at: *International Conference on Pattern Recognition and Artificial Intelligence*; August 15-17, 2018; Union, NJ, USA. [doi: [10.1145/3243250.3243268](https://doi.org/10.1145/3243250.3243268)]
19. Wei J, Huang C, Vosoughi S, Cheng Y, Xu S. Few-shot text classification with triplet networks, data augmentation, and curriculum learning. 2021 Presented at: *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*; June 6-11, 2021; Online. [doi: [10.18653/v1/2021.naacl-main.434](https://doi.org/10.18653/v1/2021.naacl-main.434)]
20. Liu H, Zhang F, Zhang X, Zhao S, Ma F, Wu XM, et al. Boosting few-shot text classification via distribution estimation. arXiv. Preprint posted online March 26, 2023. [FREE Full text]
21. Wang J, Wang KC, Rudzicz F, Brudno M. Grad2task: Improved few-shot text classification using gradients for task representation. arXiv. Preprint posted online January 27, 2022. [FREE Full text]
22. Müller M, Pérez-Torró G, Franco-Salvador M. Few-shot learning with Siamese networks and label tuning. 2022 Presented at: *The 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*; May 22-27, 2022; Dublin, Ireland. [doi: [10.18653/v1/2022.acl-long.584](https://doi.org/10.18653/v1/2022.acl-long.584)]
23. Johnson AE, Pollard TJ, Shen L, Lehman LH, Feng M, Ghassemi M, et al. MIMIC-III, a freely accessible critical care database. *Sci Data* 2016 May 24;3(1):160035 [FREE Full text] [doi: [10.1038/sdata.2016.35](https://doi.org/10.1038/sdata.2016.35)] [Medline: [27219127](https://pubmed.ncbi.nlm.nih.gov/27219127/)]
24. Sivarakumar S, Wang Y. HealthPrompt: A Zero-shot Learning Paradigm for Clinical Natural Language Processing. arXiv. Preprint posted online March 9, 2022. [FREE Full text]
25. Sentence transformers. GitHub. URL: <https://github.com/UKPLab/sentence-transformers> [accessed 2022-10-23]
26. 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 July 6, 2019. [FREE Full text]
27. Bengio Y, Courville A, Vincent P. Representation Learning: A Review and New Perspectives. *IEEE Trans. Pattern Anal. Mach. Intell* 2013 Aug;35(8):1798-1828. [doi: [10.1109/tpami.2013.50](https://doi.org/10.1109/tpami.2013.50)]
28. Fix E, Hodges JL. Discriminatory Analysis. *Nonparametric Discrimination: Consistency Properties. International Statistical Review / Revue Internationale de Statistique* 1989 Dec;57(3):238. [doi: [10.2307/1403797](https://doi.org/10.2307/1403797)]
29. SNN for FSL. GitHub. URL: <https://github.com/oniani/snn-for-fsl> [accessed 2023-04-28]
30. Good IJ. Rational Decisions. *Journal of the Royal Statistical Society: Series B (Methodological)* 2018 Dec 05;14(1):107-114. [doi: [10.1111/j.2517-6161.1952.tb00104.x](https://doi.org/10.1111/j.2517-6161.1952.tb00104.x)]
31. Loshchilov I, Hutter F. Decoupled weight decay regularization. arXiv. Preprint posted online November 14, 2017. [FREE Full text]
32. Zhang T, Wu F, Katiyar A, Weinberger KQ, Artzi Y. Revisiting few-sample BERT fine-tuning. arXiv. Preprint posted online June 10, 2020. [FREE Full text]
33. Radford A, Wu J, Child R, Luan D, Amodei D, Sutskever I. Language models are unsupervised multitask learners. *Papers with Code*. URL: <https://paperswithcode.com/paper/language-models-are-unsupervised-multitask> [accessed 2023-04-28]
34. Wolf T, Debut L, Sanh V, Chaumond J, Delangue D, Moi A, et al. Transformers: State-of-the-art natural language processing. 2020 Presented at: *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*; November 16-20, 2020; Online. [doi: [10.18653/v1/2020.emnlp-demos.6](https://doi.org/10.18653/v1/2020.emnlp-demos.6)]
35. Ge Y, Guo Y, Yang YC, Al-Garadi MA, Sarker A. Few-shot learning for medical text: A systematic review. arXiv. Preprint posted online April 21, 2022. [FREE Full text]

Abbreviations

AI: artificial intelligence

BERT: bidirectional encoder representations from transformers

BioBERT: bidirectional encoder representations from transformers for biomedical text mining
BioClinicalBERT: Bio + clinical bidirectional encoder representations from transformers
DNN: deep neural networks
FSL: few-shot learning
GPT-2: generative pretrained transformer 2
MIMIC-III: Medical Information Mart for Intensive Care
NLP: natural language processing
PLM: pretrained language model
PT-SNN: pretrained Siamese neural network
RNN: recurrent neural network
SNN: Siamese neural network
SOE-SNN: Siamese neural network with second-order embeddings

Edited by K El Emam; submitted 14.11.22; peer-reviewed by J Shi, K Koshechkin, J Zheng, A Mitra, PP Zhao; comments to author 06.02.23; revised version received 02.04.23; accepted 22.04.23; published 04.05.23.

Please cite as:

Oniani D, Chandrasekar P, Sivarajkumar S, Wang Y

Few-Shot Learning for Clinical Natural Language Processing Using Siamese Neural Networks: Algorithm Development and Validation Study

JMIR AI 2023;2:e44293

URL: <https://ai.jmir.org/2023/1/e44293>

doi: [10.2196/44293](https://doi.org/10.2196/44293)

PMID: [38875537](https://pubmed.ncbi.nlm.nih.gov/38875537/)

©David Oniani, Premkumar Chandrasekar, Sonish Sivarajkumar, Yanshan Wang. Originally published in JMIR AI (<https://ai.jmir.org>), 04.05.2023. This is an open-access article distributed under the terms of the Creative Commons Attribution License (<https://creativecommons.org/licenses/by/4.0/>), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR AI, is properly cited. The complete bibliographic information, a link to the original publication on <https://www.ai.jmir.org/>, as well as this copyright and license information must be included.

Original Paper

Extraction of Radiological Characteristics From Free-Text Imaging Reports Using Natural Language Processing Among Patients With Ischemic and Hemorrhagic Stroke: Algorithm Development and Validation

Enshuo Hsu^{1,2}, BSc, MA; Abdulaziz T Bako¹, MBBS, MPH, PhD; Thomas Potter¹, BSc, MSc, PhD; Alan P Pan¹, BSc, MSc; Gavin W Britz^{3,4}, MBBCHIR, MPH, MBA; Jonika Tannous¹, BA, PhD; Farhaan S Vahidy^{1,3,5}, MBBS, MPH, PhD

¹Center for Health Data Science and Analytics, Houston Methodist Research Institute, Houston, TX, United States

²School of Biomedical Informatics, University of Texas Health Science Center at Houston, Houston, TX, United States

³Department of Neurosurgery, Houston Methodist Neurological Institute, Houston, TX, United States

⁴Department of Neurology, Weill Cornell Medical College, New York, NY, United States

⁵Department of Population Health Sciences, Weill Cornell Medical College, New York, NY, United States

Corresponding Author:

Farhaan S Vahidy, MBBS, MPH, PhD

Center for Health Data Science and Analytics

Houston Methodist Research Institute

7550 Greenbriar Drive

Houston, TX, 77030

United States

Phone: 1 346 356 1479

Email: fvahidy@houstonmethodist.org

Abstract

Background: Neuroimaging is the gold-standard diagnostic modality for all patients suspected of stroke. However, the unstructured nature of imaging reports remains a major challenge to extracting useful information from electronic health records systems. Despite the increasing adoption of natural language processing (NLP) for radiology reports, information extraction for many stroke imaging features has not been systematically evaluated.

Objective: In this study, we propose an NLP pipeline, which adopts the state-of-the-art ClinicalBERT model with domain-specific pretraining and task-oriented fine-tuning to extract 13 stroke features from head computed tomography imaging notes.

Methods: We used the model to generate structured data sets with information on the presence or absence of common stroke features for 24,924 patients with strokes. We compared the survival characteristics of patients with and without features of severe stroke (eg, midline shift, perihematomal edema, or mass effect) using the Kaplan-Meier curve and log-rank tests.

Results: Pretrained on 82,073 head computed tomography notes with 13.7 million words and fine-tuned on 200 annotated notes, our HeadCT_BERT model achieved an average area under receiver operating characteristic curve of 0.9831, F_1 -score of 0.8683, and accuracy of 97%. Among patients with acute ischemic stroke, admissions with any severe stroke feature in initial imaging notes were associated with a lower probability of survival ($P < .001$).

Conclusions: Our proposed NLP pipeline achieved high performance and has the potential to improve medical research and patient safety.

(JMIR AI 2023;2:e42884) doi:[10.2196/42884](https://doi.org/10.2196/42884)

KEYWORDS

natural language processing; deep learning; electronic health records; ischemic stroke; cerebral hemorrhage; neuroimaging; computed tomography; stroke; radiology

Introduction

Overview

Computed tomography (CT) and magnetic resonance imaging (MRI) are the gold standards for assessing and triaging patients with suspected strokes. However, free-text imaging reports containing important radiological findings are embedded in electronic health records (EHRs) systems in an unstructured narrative format, precluding data encoding [1] to enable clinical decisions and support research applications [2-4]. Fortunately, the limitations of unstructured data have been mitigated by recent advancements in information extraction and processing methods, such as natural language processing (NLP).

Traditional rule-based NLP algorithms that use handcrafted dictionaries, keywords, and decision rules to analyze the structure of the language have classically been adopted for analyses of textual data [5-7]. However, the creation and maintenance of decision rules are labor-intensive tasks, and the quality of rules significantly influences model performance. In recent years, data-driven methods, including machine learning and deep learning, have been developed. Machine learning approaches use derived features (eg, term frequency and n-gram) from text to train supervised-learning models (eg, support vector machine [SVM] or random forest) and predict desirable outputs on new documents [3,8,9]. Deep learning methods often involve more sophisticated architectures (eg, recurrent neural networks, convolutional neural networks, and self-attention) and use word embeddings to account for the sequence and context of natural language [1,10,11].

The Bidirectional Encoder Representations from Transformers (BERT) NLP model, which uses a 24-layered deep learning architecture, was published in 2018 and achieved state-of-the-art performance on NLP benchmarks [12]. A clinical version, ClinicalBERT, was later developed by pretraining the BERT model on EHR notes to achieve improved performance on clinical data [13]. Furthermore, the ClinicalBERT model has also been trained and validated for the extraction of radiological features from chest and bone x-ray notes [14,15].

In the context of cerebrovascular disease and stroke, NLP has been applied to classify various stroke phenotypes [3,8,9] and perform feature extraction [1,5,6]. Despite these emerging applications, optimal use of NLP pipelines for stroke research is yet to be achieved. More specifically, limited studies have used BERT to extract important neuroimaging findings, such as midline shift [16] and mass effect [17]. Therefore, the use of NLP-based extraction of many critically important neuroimaging features has not been systematically implemented. We evaluated a deep learning-based NLP model (HeadCT_BERT) that is built upon ClinicalBERT and fine-tuned for the extraction and structured data generation of 13 critical stroke neuroimaging features.

Related Work

NLP on Stroke Imaging Notes

NLP has been adopted to automate stroke acuity classification. Li et al [8] used head CT and MRI radiology reports to train a

random forest model for ischemic stroke acuity classification. Kim et al [9] evaluated logistic regression, naïve Bayesian, decision tree, and SVM models to identify ischemic stroke from MRI reports. In addition, Garg et al [3] trained a variety of machine learning algorithms (ie, k-nearest neighbors, SVM, random forest, extra trees classifier, and XGBoost) to identify ischemic stroke subtypes from neurology progress notes and neuroradiology reports. In addition to NLP-based classification algorithms, a few studies adopted NLP for stroke imaging feature extraction. Yu et al [5] used a rule-based NLP tool, CHARTextract, to extract the type of occlusion, presence of established ischemia, and hemorrhage from CT reports. Gordon et al [17] proposed a machine learning-based method using XGBoost to extract the intracranial mass effect. However, there are several untapped avenues for the applications of state-of-the-art NLP methods in the stroke and cerebrovascular disease domain.

Fine-Tuning BERT for Medical Imaging Findings Extraction

The most common application of BERT is to fine-tune the out-of-box network for the NLP task. Olthof et al [18] fine-tuned the BERT model with 3268 labeled radiology reports of injured extremities and chest radiographs for extracting the presence of injury. The BERT network was appended with a binary classifier layer and trained (“fine-tuned”) with the labeled reports. The authors reported that BERT outperformed rule-based classifiers and machine learning classifiers and achieved an F_1 -score of 0.95 and an area under receiver operating characteristic curve (AUROC) of 0.99. Fink et al [19] fine-tuned the German-language BERT with structured oncology reports for rapid tumor response category classification. The results showed that the BERT model ($F_1=0.70$) achieved a similar performance as that of medical students ($F_1\approx 0.73$), although it was inferior to radiologists’ performance ($F_1=0.79$).

Pretraining and Fine-Tuning BERT for Medical Imaging Findings Extraction

Pretraining BERT with domain-specific text is an additional step that may boost model performance in subsequent fine-tuning. Smit et al [14] used an automatic labeling algorithm to tag 200,000 radiology reports for pretraining. After pretraining, 1000 reports were randomly sampled and annotated by radiologists for fine-tuning. The final NLP model, CheXbert, achieved state-of-the-art performance on one of the largest chest x-ray data sets, MIMIC-CXR, with an F_1 -score of 0.798, which is close to radiologists’ performances ($F_1=0.805$). Dai et al [15] took a similar approach using x-ray radiology reports for bone fracture. The authors developed a rule-based automatic labeling algorithm to label 6048 reports for model pretraining. Subsequently, the model was fine-tuned with a subset of 4890 manually annotated reports for fracture status detection (ie, positive, negative, or uncertain) and fracture type, bone type, and location extraction. To our knowledge, BERT pretraining in the biomedical field is underused and has not been attempted within the cerebrovascular disease domain.

Methods

Data Source and Variables

Registry for Neurological Endpoint Assessments among Patients with Ischemic and Hemorrhagic Stroke (REINAH) [20] is a data warehouse built upon the EHR at Houston Methodist, a tertiary health care system serving the greater Houston metropolitan area. REINAH hosts data for over 45,000 patients with cerebrovascular disease, representing over 982,000 neuroimaging records obtained between September 2007 and August 2022. From REINAH, we queried records that (1) had final results available before data collection on July 19, 2021; (2) had an imaging type of “CT head without contrast”; and (3) had attached imaging notes. All imaging notes were written in short paragraphs and stored as plain text. The age, sex, race, ethnicity, BMI, insurance type, stroke type, and National Institutes of Health Stroke Scale scores were extracted from each patient’s initial stroke encounter.

Ethics Approval

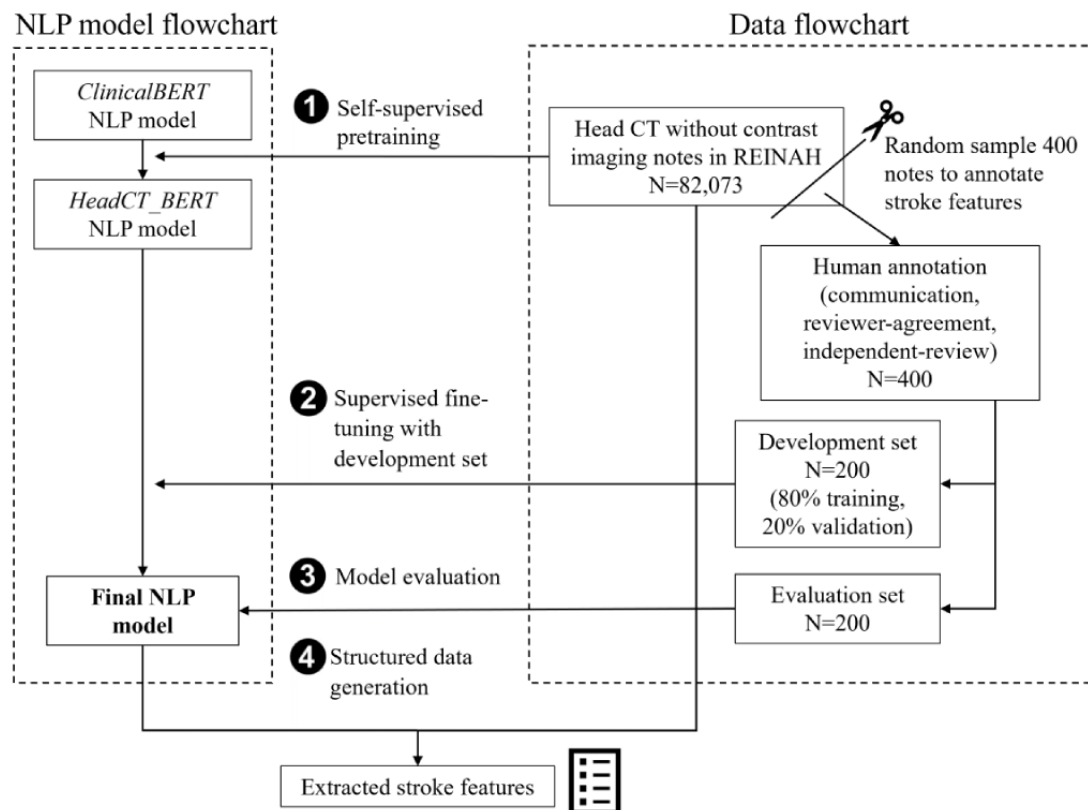
This study was approved by the Houston Methodist Institutional Review Board (PRO00025034).

Annotation

We identified 20 clinically relevant stroke-related features to extract, including hemorrhage volume, midline shift, herniation,

perihematomal edema, white matter hyperintensity, intracerebral hemorrhage (ICH) location, lacunes, old stroke, remote stroke, subacute infarct, cerebral atrophy, intraventricular hemorrhage, acute ischemia, subdural hematoma, subarachnoid hemorrhage, extra-axial hemorrhage, encephalomalacia, mass effect, and location for any non-ICH lesion (finding location). Each imaging note could include none, one, or multiple concepts. As illustrated in Figure 1, we randomly sampled 400 notes for model fine-tuning and evaluation and adopted the Begin-Inside-Outside method [21], which tags the starting position and end position of predetermined imaging features of interest in the text. We then randomly partitioned the 400 samples into the following three data sets: (1) a communication set containing 50 notes; (2) a reviewer-agreement set with 50 notes; and (3) two independent-review sets, each containing 150 notes. Two clinically trained reviewers in neuroimaging (ATB and TP) then manually annotated the imaging notes in 3 sequential stages. In the first stage, the communication set was annotated collaboratively by the 2 reviewers. In the second stage, reviewers performed separate annotations of the reviewer-agreement set, and Kappa statistics and percent agreement were evaluated. Inconsistent annotations were discussed to reach a consensus. Finally, independent review sets were separately annotated. Stroke imaging features that were identified in less than 20 notes were excluded from modeling.

Figure 1. Methodology flowchart. We used unannotated computed tomography (CT) imaging notes to pretrain the natural language processing (NLP) model and used a subset of annotated imaging notes to fine-tune and evaluate it. BERT: bidirectional encoder representations from transformers; REINAH: Registry for Neurological Endpoint Assessments among Patients with Ischemic and Hemorrhagic Stroke.



Text Processing

Before a sequence of human language can be processed by NLP models, the text often goes through processes of segmentation, tokenization, and word embedding [22]. To segment notes, we first fixed a segment length of 32 words and a step size of 10 words. For each note, the first 32 words were taken as a segment, which was then shifted to the right by 1 step (10 words) to isolate the next segment of 32 words. This process was repeated until the end of the note was reached, thereby transforming a single long note into multiple short, overlapping, text segments.

Table 1. Examples of text segmentation and word embedding^a.

Input word	Word-token(s)	Word embedding ID(s)
stroke	stroke	6625
patient	patient	5351
edema	(ed, ##ema)	(5048, 14494)
hemorrhage	(hem, ##or, ##r, ##hage)	(23123, 1766, 1197, 19911)

^aThe WordPiece algorithm takes each word as input. If a word matches a predefined word-token, embedding is done by assigning a token ID to the word. If a word does not match any predefined token, the word is split into multiple fractions and matched with predefined tokens.

Deep Learning NLP Models

Our NLP model training involved two phases, as follows: (1) an optional general training phase (“pretraining”) that familiarized the model with clinical terminology in head CT notes, and (2) a required task-specific training phase (“fine-tuning”), where the model learned to identify the 13 remaining stroke features (Table S1 in [Multimedia Appendix 1](#)).

Pretraining

Though NLP models can be trained with solely fine-tuning, recent studies have reported an improved performance after general [12,24] and domain-specific [13,25] pretraining. We used the ClinicalBERT model, which has been pretrained on general English corpora and EHR narratives [13]. We hypothesized that further pretraining it with our head CT notes using masked language model (MLM) [12] would boost the performance for stroke feature extraction. Details of NLP model pretraining are provided in Table S2 in [Multimedia Appendix 1](#). MLM used a “self-supervised” algorithm that generated labels without human annotation. A note was first tokenized into a sequence of word-tokens, and 15% of the tokens were randomly selected. Among each selected token, there was an 80% probability it would be masked (replaced by a “[MASK]” token), a 10% probability it would be replaced by a random token, and a 10% probability it remains unchanged. The MLM pretraining trained the NLP model to do “cloze,” that is, input a sequence of word-tokens with masked tokens and predict the masked tokens using the context. It is hypothesized that through learning the cloze task, the NLP model can generalize this knowledge to improve the performance of other NLP tasks. We continuously pretrained the ClinicalBERT model with 74.0k head CT imaging notes from 2007 to 2020, including a total of 13.7 million words for 5 rounds (“epochs”), and used stand-alone 8.2k notes from January to July 2021 for MLM evaluation (Table S3 in [Multimedia Appendix 1](#)). This

For each segment, word tokenization, which transforms sentences and phrases into individual word-tokens, was performed using the WordPiece [23] algorithm implemented in the Python Transformers module (version 4.10.0) and based on a predefined dictionary. In-dictionary words with predetermined tokens (eg, “stroke” and “patient”) were mapped to respective numeric IDs (word embedding). Conversely, out-of-dictionary words (eg, “edema” and “hemorrhage”) were split into multiple in-dictionary tokens and mapped to multiple token IDs (Table 1).

pretraining process produced a BERT model, which we labeled “HeadCT_BERT,” that is specific to the head CT imaging domain and can be further fine-tuned for downstream NLP tasks.

Fine-Tuning

To train the HeadCT_BERT for stroke features extraction, our downstream task in this study, we fine-tuned it with a development set of 200 notes annotated with stroke features. The HeadCT_BERT was appended with a feedforward layer with sigmoid activation function (“classification layer”) for the stroke feature classification. For each input segment (coded as a sequence of word-tokens with a maximum length of 64), the network outputs an array of probabilities (one probability for each stroke feature). The entire network (HeadCT_BERT + classification layer) was trained simultaneously. To prevent the model from becoming too attuned to the details of the development set, and consequently losing flexibility for new data (ie, to avoid overfitting), the development set was divided into a training set (80% of the notes) and a validation set (the remaining 20% of notes) [26]. Model weights were saved as checkpoints after each epoch, and optimal checkpoint weights were selected during validation as our final NLP model. The same fine-tuning process was also performed on the out-of-box ClinicalBERT model for comparison. The deep learning model was implemented using Python 3.9.6, PyTorch 1.9.0, and Transformers 4.10.0. Model computations were performed on an NVIDIA RTX 5000 graphics processing unit.

Prediction and Evaluation

The NLP model predicts the probabilities of stroke features in each segment. We aggregated the prediction to note level by selecting the maximum probability of each stroke feature among segments. The final prediction for each note consists of a probability per stroke feature (multilabel classification). We considered stroke features with a probability >.5 as presence.

To evaluate our NLP model performance, we used a stand-alone evaluation set of 200 annotated imaging notes. Evaluation metrics included recall (sensitivity), specificity, precision (positive predictive value), and F_1 -score (the harmonic mean of precision and recall). F_1 -score ranges from 0 to 1, with 1 implying perfect model performance, AUROC curve, and accuracy. We also calculated predicted probabilities and fraction of stroke features and presented probability calibration curves (reliability diagrams).



Sensitivity Analysis

One challenge for NLP modeling is the need for a large amount of human annotation, which is time consuming and labor intensive. To explore the relationship between the number of annotated training notes and model performance, and potentially reduce the annotation workload, we performed a sensitivity analysis that compared NLP models that were fine-tuned with different development set sizes: 25, 50, 100, and 150 notes. Each subset was split into a training set (80%) and a validation set (20%) and was evaluated on the set of 200 notes.

Structured Data Generation

Upon achieving satisfactory evaluation, we ran the model on all head CT imaging notes to automatically generate a structured data set of stroke imaging features. Each feature was represented as a binary variable (yes/no) associated with an imaging note. We further performed survival analysis with the Kaplan-Meier curves to evaluate the association between having any of the severe stroke features (eg, midline shift, perihematomal edema, and mass effect), as captured by NLP, and mortality for patients with acute ischemic stroke (AIS) and ICH. Differences in survival curves were compared using log-rank tests. We calculated survival rates and median survival days.

Results

Of the 982,536 available images in REINAH, we identified 82,073 head CT imaging notes representing 24,924 unique patients, of whom, 13,439 (53.9%) were female, 14,028 (56.3%) were non-Hispanic White, and 15,121 (60.7%) were Medicare beneficiaries, with an overall median age of 69 (IQR 58.5-78.3) years. With regard to stroke subtypes (at the initial encounter), 12,623 (54.4%) of patients had AIS diagnosis, 1307 (5.6%) had subarachnoid hemorrhage (SAH), 7084 (30.5%) had a transient ischemic attack (TIA), and 2208 (9.5%) had ICH. For patients with AIS, the median National Institutes of Health Stroke Scale within 6 and 12 hours of admission was 3.0 (IQR 1.0-7.0), whereas it was 7.0 (IQR 2.0, 19.0) for patients with ICH. The 400 randomly sampled notes represented 398 unique patients. Their sociodemographic characteristics were consistent with the overall population of patients with head CT images. However, a greater proportion of sampled (vs full cohort) patients had a subarachnoid hemorrhage or an ICH, perhaps owing to head CT being a gold standard for evaluation of ICH. Although median BMI was not significantly different in the annotation sample (vs full cohort), the full cohort had a

significantly higher proportion of missing BMI information (Table 2).

After annotation, stroke imaging features, including hemorrhage volume, herniation, ICH location, location of other relevant findings, remote stroke, subdural hematoma, and extra-axial hemorrhage, were excluded from modeling due to low frequencies (Table S1 in Multimedia Appendix 1). The interreviewer agreement analysis showed an excellent agreement between the 2 annotators (0.85 % average Kappa and 97.1% agreement).

Our fine-tuned HeadCT_BERT model had an AUROC of 0.9831 and an F_1 -score of 0.8683. The F_1 -scores were greater than 0.9 for 8 of 13 (61.5%) stroke imaging features, and the AUROCs were greater than 0.96 for all features except for acute ischemia. Results show that after fine-tuning, both ClinicalBERT and HeadCT_BERT achieved favorable performances, while HeadCT_BERT demonstrated marginally better performance (Table 3 and Table 4; Figure S2 in Multimedia Appendix 1).

The sensitivity analysis revealed sigmoid shapes for both models, indicating that improvement in model performance wanes as sample size approaches an optimal point. Specifically, we found marked performance improvements when increasing the training sample size from 25 to 50 and 100 notes. From 100 to 150, however, performance gain decreases, and from 150 to 200 notes, the performance gain is minimal, indicating that the NLP models had achieved near-optimal performance (Figure S1 in Multimedia Appendix 1).

The probability calibration curves showed HeadCT_BERT is well calibrated for some stroke features (eg, midline shift, white matter hyperintensity, subacute infarct, acute ischemia, subarachnoid hemorrhage, and encephalomalacia), while ClinicalBERT is well calibrated for midline shift, white matter hyperintensity, old stroke, subacute infarct, cerebral atrophy, acute ischemia, ICH, encephalomalacia, and mass effect (Figure S3 in Multimedia Appendix 1).

Running on a single-graphics processing unit server, our final NLP model processed ~230 imaging notes per minute and automatically generated a structured stroke imaging feature data set from 24,924 patients with head CT notes across the hospital system. In the resulting data set, 3826 (15.4%) of patients had a mass effect, 3600 (14.4%) had perihematomal edema, 1908 (7.7%) had a midline shift, and 5146 (20.6%) had 1 or more than 1 severe stroke features (eg, midline shift, mass effect, or perihematomal edema; Table 5).

Survival analysis based on the initial head CT notes of 6463 AIS and 1243 ICH emergency admissions showed that patients with severe stroke features had higher mortality and shorter survival times (AIS: 18.4% mortality rate and 585 days median survival time; ICH: 20.7% mortality rate and 572 days median survival time) compared to other patients (AIS: 10.1% mortality rate and 759 days median survival time; ICH: 17.8% mortality rate and 638 days median survival time). Differences in survival probability over time are shown as Kaplan-Meier curves. Among AIS admissions, patients with severe stroke features had significantly lower survival probabilities ($P<.001$; Figure 2).

Table 2. Patient characteristics (average age and BMI are reported at imaging encounters). Italicized *P* values are significant.

Characteristics	Head CT ^a population	Annotation sample	<i>P</i> value
Imaging notes, N	82,073	400	
Unique patients, N	24,924	398	
Age (years), median (Q1, Q3)	69.0 (58.5, 78.3)	68.0 (56.4, 78.1)	.22
Age (years), n (%)			.41
0-49	3025 (12.1)	57 (14.3)	
50-59	3793 (15.2)	61 (15.3)	
60-69	6149 (24.7)	103 (25.9)	
≥70	11,957 (48)	177 (44.5)	
Gender, n (%)			.69
Female	13,439 (53.9)	219 (55)	
Male	11,485 (46.1)	179 (45)	
Race or ethnicity, n (%)			.22
Non-Hispanic White	14,028 (56.3)	206 (51.8)	
Black	5690 (22.8)	102 (25.6)	
Hispanic	3412 (13.7)	61 (15.3)	
Asian	1209 (4.9)	16 (4)	
Other or unknown	585 (2.3)	13 (3.3)	
BMI (kg/m ²), median (Q1, Q3)	27.3 (23.7, 31.7)	27.3 (23.5, 31.0)	.59
BMI (kg/m²), n (%)			.001
Underweight	637 (2.6)	13 (3.3)	
Normal	6193 (24.8)	108 (27.1)	
Overweight	6518 (26.2)	123 (30.9)	
Obese	6610 (26.5)	107 (26.9)	
Missing	4966 (19.9)	47 (11.8)	
Insurance^b, n (%)			
Medicare			.15
No	9803 (39.3)	142 (35.7)	
Yes	15,121 (60.7)	256 (64.3)	
Medicaid			.12
No	23,793 (95.5)	373 (93.7)	
Yes	1131 (4.5)	25 (6.3)	
Commercial			.04
No	20,194 (81)	306 (76.9)	
Yes	4730 (19)	92 (23.1)	
Exchange			.79
No	24,437 (98)	389 (97.7)	
Yes	487 (2)	9 (2.3)	
Primary stroke type^c, n (%)			<.001
Subarachnoid hemorrhage	1307 (5.6)	29 (7.7)	
Transient ischemic attack	7084 (30.5)	100 (26.5)	
Intracerebral hemorrhage	2208 (9.5)	59 (15.6)	

Characteristics	Head CT ^a population	Annotation sample	P value
Acute ischemic stroke	12,623 (54.4)	189 (50.1)	
NIHSS^d Stroke Scale for acute ischemic stroke, median (Q1, Q3)			
Average NIHSS in 6 hours	3.0 (1.0, 7.0)	3.0 (1.5, 9.0)	.09
Average NIHSS in 12 hours	3.0 (1.0, 7.0)	3.0 (1.0, 8.0)	.24
NIHSS Stroke Scale for intracerebral hemorrhage, median (Q1, Q3)			
Average NIHSS in 6 hours	7.0 (2.0, 19.0)	6 (1.5, 18.0)	.94
Average NIHSS in 12 hours	7.0 (2.0, 19.0)	7.0 (2.0, 18.0)	.81

^aCT: computed tomography.

^bInsurance type was collected throughout all imaging encounters.

^cFor patients with multiple stroke visits, the initial encounter's stroke scale and primary stroke type are presented. We perform hypothesis testing to compare the 398 sampled patients with the nonsampled population. Chi-square tests were adopted for categorical variables, and Kruskal-Wallis tests were adopted for continuous variables.

^dNIHSS: National Institutes of Health Stroke Scale.

Table 3. Final natural language processing model evaluation with the evaluation set (N=200) at the imaging note level.

Stroke feature	Specificity	Precision	Recall	F ₁ -score	AUROC ^a (95% CI)	Accuracy (95% CI)
Midline shift	1	1	0.9375	0.9677	0.9973 (0.9792-1.0154)	0.9950 (0.9852-1.0048)
Perihematomal edema	0.9945	0.9474	0.9474	0.9474	0.9994 (0.9917-1.0071)	0.9900 (0.9762-1.0038)
White matter hyperintensity	0.9725	0.9667	0.956	0.9613	0.9704 (0.9452-0.9955)	0.9650 (0.9395-0.9905)
Lacunae	1	1	1	1	1.0000 (1.0000-1.0000)	1.0000 (1.0000-1.0000)
Old stroke	0.9581	0.8056	0.8788	0.8406	0.9693 (0.9277-1.0110)	0.9450 (0.9134-0.9766)
Subacute infarct	0.9945	0.9091	0.5556	0.6897	0.9789 (0.9321-1.0258)	0.9550 (0.9263-0.9837)
Cerebral atrophy	0.9173	0.8571	0.9851	0.9167	0.9673 (0.9369-0.9978)	0.9400 (0.9071-0.9729)
Intraventricular hemorrhage	0.984	0.7273	0.6154	0.6667	0.9798 (0.9259-1.0338)	0.9600 (0.9328-0.9872)
Acute ischemia	0.956	0.6364	0.7778	0.7	0.9362 (0.8570-1.0154)	0.9400 (0.9071-0.9729)
Intracerebral hemorrhage	0.9665	0.75	0.8571	0.8	0.9872 (0.9532-1.0212)	0.9550 (0.9263-0.9837)
Subarachnoid hemorrhage	1	1	0.8333	0.9091	1.0000 (1.0000-1.0000)	0.9900 (0.9762-1.0038)
Encephalomalacia	1	1	0.9524	0.9756	0.9989 (0.9890-1.0088)	0.9950 (0.9852-1.0048)
Mass effect	0.9777	0.84	1	0.913	0.9952 (0.9743-1.0161)	0.9800 (0.9606-0.9994)

^aAUROC: area under receiver operating characteristic curve.

Table 4. Average natural language processing model evaluation metrics among 13 stroke features for the fine-tuned models.

Stroke feature	F ₁ -score, mean (SD)	AUROC ^a , mean (SD)	Accuracy, mean (SD)
HeadCT_BERT (final model)	<i>0.8683 (0.1176)^b</i>	<i>0.9831 (0.0189)^b</i>	<i>0.9700 (0.0225)^b</i>
ClinicalBERT (baseline model)	0.8564 (0.1173)	0.9786 (0.0216)	0.9665 (0.0237)

^aAUROC: area under receiver operating characteristic curve.

^bItalicized values denote performance of the proposed model.

Table 5. Natural language processing (NLP) model generating structured stroke feature data sets from imaging notes^a.

Characteristics	Head CT ^b imaging patients ^c (N=24924), n (%)	Acute ischemic stroke admission initial CT ^d (N=6463), n (%)	Intracerebral hemorrhage admission initial CT ^e (N=1243), n (%)
White matter hyperintensity	16,014 (64.3)	3429 (53.1)	407 (32.7)
Cerebral atrophy	13,615 (54.6)	2262 (35)	268 (21.6)
Old stroke	7426 (29.8)	1324 (20.5)	91 (7.3)
Lacunae	6622 (26.6)	1386 (21.4)	116 (9.3)
Mass effect	3826 (15.4)	614 (9.5)	500 (40.2)
Intracerebral hemorrhage	3822 (15.3)	354 (5.5)	1096 (88.2)
Perihematomal edema	3600 (14.4)	436 (6.7)	623 (50.1)
Encephalomalacia	3453 (13.9)	373 (5.8)	50 (4)
Acute ischemia	3426 (13.7)	1173 (18.1)	33 (2.7)
Subacute infarct	2675 (10.7)	841 (13)	28 (2.3)
subarachnoid hemorrhage	2179 (8.7)	132 (2)	245 (19.7)
Midline shift	1908 (7.7)	184 (2.8)	345 (27.8)
Intraventricular hemorrhage	1409 (5.7)	37 (0.6)	405 (32.6)
Severe stroke features ^f	5146 (20.6)	901 (13.9)	845 (68)

^aOur final NLP model processed 82,073 head computed tomography notes for 24,924 unique patients in the entire hospital system and generated structured data sets.

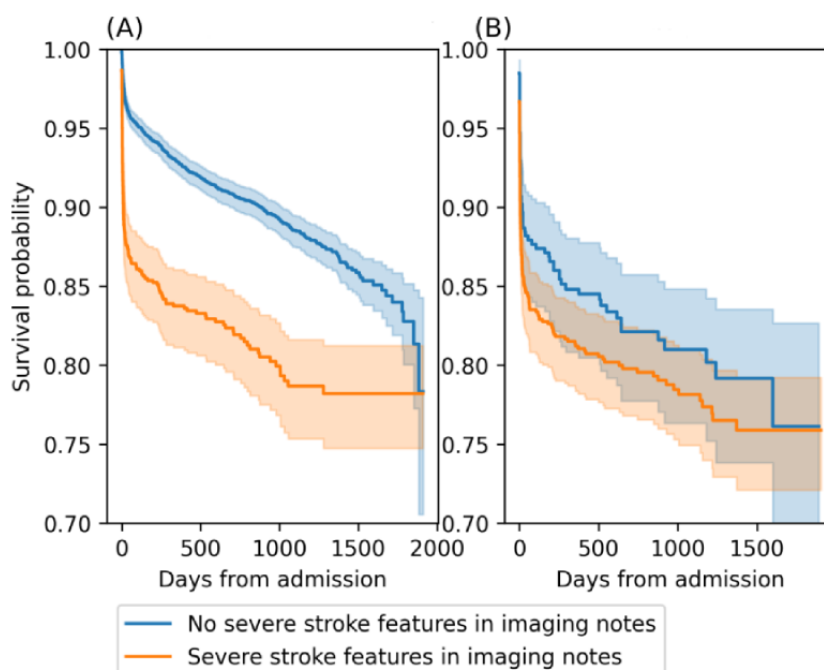
^bCT: computed tomography.

^cThe stroke features in the overall population were aggregated at the patient level.

^{d,e}The stroke features in the initial head CT of acute ischemic stroke and intracerebral hemorrhage emergency admissions were presented.

^fSevere stroke features include midline shift, perihematomal edema, or mass effect. Severe stroke feature is a composite feature.

Figure 2. Kaplan-Meier curve of survival probability from initial admissions. Patients whose initial imaging includes severe stroke features (eg, midline shift, mass effect, or perihematomal edema) had a lower survival probability. (A) Acute ischemic stroke admissions ($P<.001$). (B) Intracerebral hemorrhage admissions ($P=.19$).



Discussion

Principal Findings

We propose an NLP pipeline to extract ischemic and hemorrhagic stroke characteristics from head CT imaging notes (HeadCT_BERT model). Built upon one of the latest clinical NLP models, the HeadCT_BERT model achieved an excellent average AUROC of 0.9831 and an accuracy of 97%. Our NLP pipeline showed promising performance for the detection of midline shift, perihematomal edema, lacunes, subarachnoid hemorrhage, encephalomalacia, and mass effect, with AUROCs for each of these features exceeding 0.99 and F_1 -scores above 0.9 for the evaluation set. Other features, including white matter hyperintensity, old stroke, subacute infarct, cerebral atrophy, intraventricular hemorrhage, and ICH showed AUROCs between 0.96 to 0.98. Other NLP studies have achieved optimal AUROC values of 0.9625 for mass effect extraction [17], 0.96 for stroke presence, and 0.93 for stroke acuity [1]. Our method achieved comparable or better performance for extracting stroke imaging features.

In 2018 alone, 11.5 million head CT scans were performed in the United States [27], generating valuable information that can be used to answer a multitude of stroke-related research questions. In the absence of methods to extract information in unstructured formats, the generation of insights from such sources is limited. This underscores the value of our NLP pipeline, which provides a fast, scalable, and automatic solution for the processing of unstructured text data.

Application of our pipeline in a health care environment has the potential to benefit both medical research and patient safety. For example, in this study, we demonstrated the use of NLP for retrospectively identifying cohorts of patients with AIS and ICH with severe stroke features. We identified 901 (13.9%) AIS and 845 (68%) patients with ICH with severe stroke neuroimaging features and demonstrated lower survival rates for patients with these severe features, consistent with previous studies [28,29]. Beyond outcome prediction, modifications of our pipeline may also be implemented to improve patient safety. For example, NLP pipelines that detect incidents can be used

to improve patient outreach workflows by optimizing reporting procedures for health care providers as well as the patients and their families [30]. Our pipeline has the potential to process imaging notes in real time, generate flags for severe stroke findings, and trigger reminders and alerts within the EHR system.

Despite the performance of our NLP pipeline, this study has limitations. First, it was conducted and evaluated in a single organization, where many of the notes may have been written by a relatively small number of radiologists or neuroradiologists. Therefore, the generalizability of the trained NLP models could be limited by overly consistent wording and grammar in training data. However, as one of the largest hospital systems, comprising 7 certified stroke care hospitals in the Houston metropolitan area, we feel that our inclusion of a diverse collection of notes yields enough variability in the training data to mitigate this issue. Second, although our HeadCT_BERT model demonstrated slightly improved performance for stroke features extraction, it is hard to compare our model with ClinicalBERT due to the lack of well-established NLP benchmarks for head imaging reports. Future efforts to create head imaging NLP benchmarks are needed for comprehensive evaluation. Finally, the probability calibration curves of both HeadCT_BERT and ClinicalBERT for individual stroke features demonstrate a mixed performance in calibration, indicating potential imbalance of certain stroke features in the training data set. As a result, using a probability of .5 as a general cut-off might not be optimal for all stroke features. Future work is required to adequately calibrate the model for all stroke features.

Conclusions

This study represents a step forward in NLP adoption for neuroimaging among patients with cerebrovascular disease. Our work demonstrates an effective and customizable NLP pipeline for retrieving multiple stroke features from large amounts of unstructured imaging notes. Derived from the latest artificial intelligence technology, we believe our model will benefit stroke research and patient safety. To fully understand the impact on the health care industry, future work in the data pipeline deployment and evaluation is anticipated.

Acknowledgments

This project is supported by the Center for Health Data Science and Analytics, Department of Neurosurgery, and the Neurological Institute at Houston Methodist.

Authors' Contributions

EH conceived the study and performed data analysis and Natural language processing modeling. ATB and TP helped with manual annotation. All authors contributed to the manuscript writing.

Conflicts of Interest

None declared.

Multimedia Appendix 1

Supplementary tables and figures.

[DOCX File, 898 KB - ai_v2i1e42884_app1.docx]

References

1. Ong CJ, Orfanoudaki A, Zhang R, Caprasse FPM, Hutch M, Ma L, et al. Machine learning and natural language processing methods to identify ischemic stroke, acuity and location from radiology reports. *PLoS One* 2020;15(6):e0234908 [FREE Full text] [doi: [10.1371/journal.pone.0234908](https://doi.org/10.1371/journal.pone.0234908)] [Medline: [32559211](https://pubmed.ncbi.nlm.nih.gov/32559211/)]
2. Grivas A, Alex B, Grover C, Tobin R, Whiteley W. Not a cute stroke: analysis of rule- and neural network-based information extraction systems for brain radiology reports. *Proceedings of the 11th International Workshop on Health Text Mining and Information Analysis* 2020:24-37. [doi: [10.18653/v1/2020.louhi-1.4](https://doi.org/10.18653/v1/2020.louhi-1.4)]
3. Garg R, Oh E, Naidech A, Kording K, Prabhakaran S. Automating ischemic stroke subtype classification using machine learning and natural language processing. *J Stroke Cerebrovasc Dis* 2019 Jul;28(7):2045-2051. [doi: [10.1016/j.jstrokecerebrovasdis.2019.02.004](https://doi.org/10.1016/j.jstrokecerebrovasdis.2019.02.004)] [Medline: [31103549](https://pubmed.ncbi.nlm.nih.gov/31103549/)]
4. Sorin V, Barash Y, Konen E, Klang E. Deep learning for natural language processing in radiology-fundamentals and a systematic review. *J Am Coll Radiol* 2020 May;17(5):639-648. [doi: [10.1016/j.jacr.2019.12.026](https://doi.org/10.1016/j.jacr.2019.12.026)] [Medline: [32004480](https://pubmed.ncbi.nlm.nih.gov/32004480/)]
5. Yu AXY, Liu ZA, Pou-Prom C, Lopes K, Kapral MK, Aviv RI, et al. Automating stroke data extraction from free-text radiology reports using natural language processing: instrument validation study. *JMIR Med Inform* 2021 May 04;9(5):e24381 [FREE Full text] [doi: [10.2196/24381](https://doi.org/10.2196/24381)] [Medline: [33944791](https://pubmed.ncbi.nlm.nih.gov/33944791/)]
6. Wheeler E, Mair G, Sudlow C, Alex B, Grover C, Whiteley W. A validated natural language processing algorithm for brain imaging phenotypes from radiology reports in UK electronic health records. *BMC Med Inform Decis Mak* 2019 Sep 09;19(1):184. [doi: [10.1186/s12911-019-0908-7](https://doi.org/10.1186/s12911-019-0908-7)] [Medline: [31500613](https://pubmed.ncbi.nlm.nih.gov/31500613/)]
7. Berman AN, Biery DW, Ginder C, Hulme OL, Marcusa D, Leiva O, et al. Natural language processing for the assessment of cardiovascular disease comorbidities: the Cardio-Canary comorbidity project. *Clin Cardiol* 2021 Sep;44(9):1296-1304 [FREE Full text] [doi: [10.1002/clc.23687](https://doi.org/10.1002/clc.23687)] [Medline: [34347314](https://pubmed.ncbi.nlm.nih.gov/34347314/)]
8. Li MD, Lang M, Deng F, Chang K, Buch K, Rincon S, et al. Analysis of stroke detection during the COVID-19 pandemic using natural language processing of radiology reports. *AJNR Am J Neuroradiol* 2021 Mar;42(3):429-434 [FREE Full text] [doi: [10.3174/ajnr.A6961](https://doi.org/10.3174/ajnr.A6961)] [Medline: [33334851](https://pubmed.ncbi.nlm.nih.gov/33334851/)]
9. Kim C, Zhu V, Obeid J, Lenert L. Natural language processing and machine learning algorithm to identify brain MRI reports with acute ischemic stroke. *PLoS One* 2019;14(2):e0212778 [FREE Full text] [doi: [10.1371/journal.pone.0212778](https://doi.org/10.1371/journal.pone.0212778)] [Medline: [30818342](https://pubmed.ncbi.nlm.nih.gov/30818342/)]
10. Heo TS, Kim YS, Choi JM, Jeong YS, Seo SY, Lee JH, et al. Prediction of stroke outcome using natural language processing-based machine learning of radiology report of brain MRI. *J Pers Med* 2020 Dec 16;10(4):286 [FREE Full text] [doi: [10.3390/jpm10040286](https://doi.org/10.3390/jpm10040286)] [Medline: [33339385](https://pubmed.ncbi.nlm.nih.gov/33339385/)]
11. Wood DA, Kafiabadi S, Al Busaidi A, Guilhem EL, Lynch J, Townend MK, et al. Deep learning to automate the labelling of head MRI datasets for computer vision applications. *Eur Radiol* 2022 Jan;32(1):725-736 [FREE Full text] [doi: [10.1007/s00330-021-08132-0](https://doi.org/10.1007/s00330-021-08132-0)] [Medline: [34286375](https://pubmed.ncbi.nlm.nih.gov/34286375/)]
12. Devlin J, Chang M, Lee K, Toutanova K. BERT: Pre-training of deep bidirectional transformers for language understanding. *ArXiv Preprint* posted online May 24, 2019. [FREE Full text]
13. Alsentzer E, Murphy J, Boag W, Weng W, Jin D, Naumann T, et al. Publicly available clinical BERT embeddings. *ArXiv Preprint* posted online June 6, 2019. [FREE Full text] [doi: [10.18653/v1/w19-1909](https://doi.org/10.18653/v1/w19-1909)]
14. Smit A, Jain S, Rajpurkar P, Pareek A, Ng A, Lungren M. CheXbert: Combining automatic labelers and expert annotations for accurate radiology report labeling using BERT. *ArXiv Preprint* posted online Apr 20, 2022. [FREE Full text] [doi: [10.18653/v1/2020.emnlp-main.117](https://doi.org/10.18653/v1/2020.emnlp-main.117)]
15. Dai Z, Li Z, Han L. BoneBert: A BERT-based automated information extraction system of radiology reports for bone fracture detection and diagnosis. In: *IDA 2021: Advances in Intelligent Data Analysis XIX*. Cham, Switzerland: Springer, Cham; 2021:263-274.
16. Pruitt P, Naidech A, Van Ornam J, Borczuk P, Thompson W. A natural language processing algorithm to extract characteristics of subdural hematoma from head CT reports. *Emerg Radiol* 2019 Jun;26(3):301-306. [doi: [10.1007/s10140-019-01673-4](https://doi.org/10.1007/s10140-019-01673-4)] [Medline: [30693414](https://pubmed.ncbi.nlm.nih.gov/30693414/)]
17. Gordon AJ, Banerjee I, Block J, Winstead-Derlega C, Wilson JG, Mitarai T, et al. Natural language processing of head CT reports to identify intracranial mass effect: CTIME algorithm. *Am J Emerg Med* 2022 Jan;51:388-392. [doi: [10.1016/j.ajem.2021.11.001](https://doi.org/10.1016/j.ajem.2021.11.001)] [Medline: [34839182](https://pubmed.ncbi.nlm.nih.gov/34839182/)]
18. Olthof AW, Shouche P, Fennema EM, IJpma FFA, Koolstra RHC, Stirler VMA, et al. Machine learning based natural language processing of radiology reports in orthopaedic trauma. *Comput Methods Programs Biomed* 2021 Sep;208:106304 [FREE Full text] [doi: [10.1016/j.cmpb.2021.106304](https://doi.org/10.1016/j.cmpb.2021.106304)] [Medline: [34333208](https://pubmed.ncbi.nlm.nih.gov/34333208/)]
19. Fink MA, Kades K, Bischoff A, Moll M, Schnell M, Kuchler M, et al. Deep learning-based assessment of oncologic outcomes from natural language processing of structured radiology reports. *Radiol Artif Intell* 2022 Sep;4(5):e220055 [FREE Full text] [doi: [10.1148/ryai.220055](https://doi.org/10.1148/ryai.220055)] [Medline: [36204531](https://pubmed.ncbi.nlm.nih.gov/36204531/)]
20. Potter T, Pratap S, Nicolas J, Khan O, Alan P, Bako A, et al. A neuro-informatics pipeline to support a learning healthcare system for populations with cerebrovascular disease: rationale and design for a registry across an 8-hospital tertiary healthcare system in the greater Houston metropolitan area. *JMIR preprints* Preprint posted online on June 30, 2022. [doi: [10.2196/preprints.40639](https://doi.org/10.2196/preprints.40639)]

21. Ratinov L, Roth D. Design challenges and misconceptions in named entity recognition. In: Proceedings of the Thirteenth Conference on Computational Natural Language Learning (CoNLL-2009). 2009 Presented at: CoNLL '09; June 4-5; Boulder, CO p. 147-155 URL: <https://aclanthology.org/W09-1119> [doi: [10.3115/1596374.1596399](https://doi.org/10.3115/1596374.1596399)]
22. Mozayan A, Fabbri AR, Maneveese M, Tocino I, Chheang S. Practical guide to natural language processing for radiology. *Radiographics* 2021 Sep;41(5):1446-1453. [doi: [10.1148/rg.2021200113](https://doi.org/10.1148/rg.2021200113)] [Medline: [34469212](https://pubmed.ncbi.nlm.nih.gov/34469212/)]
23. Wu Y, Schuster M, Chen Z, Le Q, Norouzi M, Macherey W, et al. Google's neural machine translation system: bridging the gap between human and machine translation. *ArXiv Preprint* posted online Oct 8, 2016. [FREE Full text] [doi: [10.1162/tacl_a_00065](https://doi.org/10.1162/tacl_a_00065)]
24. 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 July 26, 2019. [FREE Full text]
25. Lee J, Yoon W, Kim S, Kim D, Kim S, So C, et al. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics* 2020 Feb 15;36(4):1234-1240 [FREE Full text] [doi: [10.1093/bioinformatics/btz682](https://doi.org/10.1093/bioinformatics/btz682)] [Medline: [31501885](https://pubmed.ncbi.nlm.nih.gov/31501885/)]
26. Xu Y, Goodacre R. On splitting training and validation set: a comparative study of cross-validation, bootstrap and systematic sampling for estimating the generalization performance of supervised learning. *J Anal Test* 2018 Oct 29;2(3):249-262 [FREE Full text] [doi: [10.1007/s41664-018-0068-2](https://doi.org/10.1007/s41664-018-0068-2)] [Medline: [30842888](https://pubmed.ncbi.nlm.nih.gov/30842888/)]
27. Cauley KA, Hu Y, Fielden SW. Head CT: toward making full use of the information the X-rays give. *AJNR Am J Neuroradiol* 2021 Aug;42(8):1362-1369 [FREE Full text] [doi: [10.3174/ajnr.A7153](https://doi.org/10.3174/ajnr.A7153)] [Medline: [34140278](https://pubmed.ncbi.nlm.nih.gov/34140278/)]
28. Nag C, Das K, Ghosh M, Khandakar MR. Prediction of clinical outcome in acute hemorrhagic stroke from a single CT Scan on admission. *N Am J Med Sci* 2012 Oct;4(10):463-467 [FREE Full text] [doi: [10.4103/1947-2714.101986](https://doi.org/10.4103/1947-2714.101986)] [Medline: [23112967](https://pubmed.ncbi.nlm.nih.gov/23112967/)]
29. Daverat P, Castel JP, Dartigues JF, Orgogozo JM. Death and functional outcome after spontaneous intracerebral hemorrhage. A prospective study of 166 cases using multivariate analysis. *Stroke* 1991 Jan;22(1):1-6. [doi: [10.1161/01.str.22.1.1](https://doi.org/10.1161/01.str.22.1.1)] [Medline: [1987664](https://pubmed.ncbi.nlm.nih.gov/1987664/)]
30. Canton SP, Dadashzadeh E, Yip L, Forsythe R, Handzel R. Automatic detection of thyroid and adrenal incidentals using radiology reports and deep learning. *J Surg Res* 2021 Oct;266:192-200. [doi: [10.1016/j.jss.2021.03.060](https://doi.org/10.1016/j.jss.2021.03.060)] [Medline: [34020097](https://pubmed.ncbi.nlm.nih.gov/34020097/)]

Abbreviations

AIS: acute ischemic stroke

AUROC: area under receiver operating characteristic curve

BERT: bidirectional encoder representations from transformers

CT: computed tomography

EHR: electronic health record

ICH: intracerebral hemorrhage

MLM: masked language model

MRI: magnetic resonance imaging

NLP: natural language processing

REINAH: Registry for Neurological Endpoint Assessments among Patients with Ischemic and Hemorrhagic Stroke

SVM: Support vector machine

Edited by K El Emam, B Malin; submitted 22.09.22; peer-reviewed by M Kapsetaki, O Balogun, W Klement; comments to author 11.11.22; revised version received 10.01.23; accepted 08.04.23; published 06.06.23.

Please cite as:

Hsu E, Bako AT, Potter T, Pan AP, Britz GW, Tannous J, Vahidy FS

Extraction of Radiological Characteristics From Free-Text Imaging Reports Using Natural Language Processing Among Patients With Ischemic and Hemorrhagic Stroke: Algorithm Development and Validation

JMIR AI 2023;2:e42884

URL: <https://ai.jmir.org/2023/1/e42884>

doi: [10.2196/42884](https://doi.org/10.2196/42884)

PMID: [38875556](https://pubmed.ncbi.nlm.nih.gov/38875556/)

©Enshuo Hsu, Abdulaziz T Bako, Thomas Potter, Alan P Pan, Gavin W Britz, Jonika Tannous, Farhaan S Vahidy. Originally published in *JMIR AI* (<https://ai.jmir.org>), 06.06.2023. This is an open-access article distributed under the terms of the Creative Commons Attribution License (<https://creativecommons.org/licenses/by/4.0/>), which permits unrestricted use, distribution, and

reproduction in any medium, provided the original work, first published in JMIR AI, is properly cited. The complete bibliographic information, a link to the original publication on <https://www.ai.jmir.org/>, as well as this copyright and license information must be included.

Original Paper

Natural Language Processing for Clinical Laboratory Data Repository Systems: Implementation and Evaluation for Respiratory Viruses

Elham Dolatabadi^{1,2,3}, BSc, MSc, PhD; Branson Chen⁴, BHS, MSc; Sarah A Buchan^{3,4,5,6}, MSc, PhD; Alex Marchand Austin⁴, BSc, MSc; Mahmoud Azimae^{3,4}, BSc; Allison McGeer^{3,6,7,8}, MSc, MD, PhD; Samira Mubareka^{8,9}, MD; Jeffrey C Kwong^{4,5,6,10,11}, MSc, MD

¹Vector Institute, Toronto, ON, Canada

²School of Health Policy and Management, Faculty of Health, York University, Toronto, ON, Canada

³Institute of Health Policy, Management and Evaluation, University of Toronto, Toronto, ON, Canada

⁴ICES, Toronto, ON, Canada

⁵Public Health Ontario, Toronto, ON, Canada

⁶Dalla Lana School of Public Health, University of Toronto, Toronto, ON, Canada

⁷Sinai Health System, Toronto, ON, Canada

⁸Department of Laboratory Medicine and Pathobiology, University of Toronto, Toronto, ON, Canada

⁹Sunnybrook Research Institute, Toronto, ON, Canada

¹⁰University Health Network, Toronto, ON, Canada

¹¹Department of Family and Community Medicine, University of Toronto, Toronto, ON, Canada

Corresponding Author:

Elham Dolatabadi, BSc, MSc, PhD

Vector Institute

661 University Ave

Toronto, ON, M5G 1M1

Canada

Phone: 1 6477069756

Email: elham.dolatabadi@gmail.com

Abstract

Background: With the growing volume and complexity of laboratory repositories, it has become tedious to parse unstructured data into structured and tabulated formats for secondary uses such as decision support, quality assurance, and outcome analysis. However, advances in natural language processing (NLP) approaches have enabled efficient and automated extraction of clinically meaningful medical concepts from unstructured reports.

Objective: In this study, we aimed to determine the feasibility of using the NLP model for information extraction as an alternative approach to a time-consuming and operationally resource-intensive handcrafted rule-based tool. Therefore, we sought to develop and evaluate a deep learning-based NLP model to derive knowledge and extract information from text-based laboratory reports sourced from a provincial laboratory repository system.

Methods: The NLP model, a hierarchical multilabel classifier, was trained on a corpus of laboratory reports covering testing for 14 different respiratory viruses and viral subtypes. The corpus includes 87,500 unique laboratory reports annotated by 8 subject matter experts (SMEs). The classification task involved assigning the laboratory reports to labels at 2 levels: 24 fine-grained labels in level 1 and 6 coarse-grained labels in level 2. A “label” also refers to the status of a specific virus or strain being tested or detected (eg, influenza A is detected). The model’s performance stability and variation were analyzed across all labels in the classification task. Additionally, the model’s generalizability was evaluated internally and externally on various test sets.

Results: Overall, the NLP model performed well on internal, out-of-time (pre-COVID-19), and external (different laboratories) test sets with microaveraged F_1 -scores >94% across all classes. Higher precision and recall scores with less variability were observed for the internal and pre-COVID-19 test sets. As expected, the model’s performance varied across categories and virus types due to the imbalanced nature of the corpus and sample sizes per class. There were intrinsically fewer classes of viruses

being detected than those tested; therefore, the model's performance (lowest F_1 -score of 57%) was noticeably lower in the detected cases.

Conclusions: We demonstrated that deep learning–based NLP models are promising solutions for information extraction from text-based laboratory reports. These approaches enable scalable, timely, and practical access to high-quality and encoded laboratory data if integrated into laboratory information system repositories.

(*JMIR AI 2023;2:e44835*) doi:[10.2196/44835](https://doi.org/10.2196/44835)

KEYWORDS

health; informatics; natural language processing; knowledge extraction; electronic health record; EHR

Introduction

Clinical laboratory data account for a large proportion of data stored in electronic health record systems worldwide and present a wealth of information vital for evidence-based decision-making and public health improvement [1,2]. Laboratory information systems record, manage, and store laboratory test data to facilitate reporting to clinicians and jurisdictional laboratory information repositories [3]. These repositories often include test orders and results from various laboratory service providers, such as hospitals, public health agencies, and private companies, and are populated as part of clinical care.

Several factors limit the secondary use of laboratory data for other purposes. The most important are concerns about the quality of the data, lack of standardization, and difficulty extracting the needed information [4,5]. Laboratory data vary over time due to evolving standards of care and changing population demographics. Furthermore, specific categories of laboratory data are reported as free text in an unstructured format with no standard vocabulary in the actual contents, which adds more complexity for their secondary uses [1]. Therefore, efforts are needed to eliminate redundancies, extract the necessary information, and derive accurate interpretations from laboratory data.

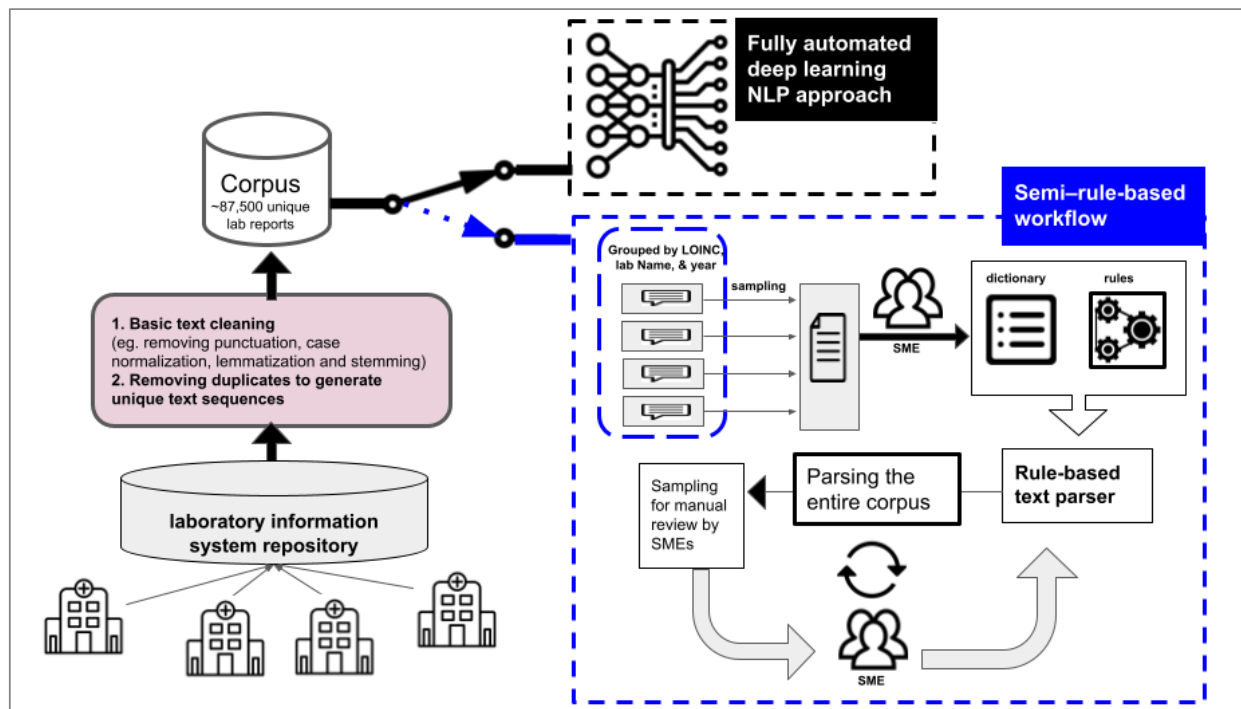
Our institute, ICES, has developed a specific information extraction workflow to manage the interpretation of a large volume of provincial clinical laboratory results, as shown in [Figure 1](#). The workflow, called a semi–rule-based workflow, relies on time-consuming and operationally resource-intensive approaches, including a library of rule-based and handcrafted tools. These tools are explicitly programmed for various laboratory result categories and must be refined continually. To address challenges with our existing semi–rule-based workflow and automate the exhaustive information retrieval task, we built a deep learning–based natural language processing (NLP) tool.

The objective of this study was to assess the feasibility of our deep learning–based NLP model and evaluate its performance relative to the semi–rule-based workflow.

The development of NLP methods is essential to automatically transform laboratory reports into a structured representation that scales data usability for research, quality improvement, and clinical purposes [6–12]. NLP enables automated extraction of information, and its use in the clinical domain is growing, with increasing uptake in various applications such as biomedical named entity recognition [11,12], summarization [10], and clinical prediction tasks [9]. More recently, deep learning approaches such as convolutional neural networks, recurrent neural networks (RNNs), and RNN variants such as bidirectional long short-term memory (Bi-LSTM) have been successfully applied to clinical NLP tasks [10,13–16]. They are now considered the baseline techniques for various information extraction tasks [11,12,17–20].

In this study, we focused on automating the retrieval of information related to respiratory viruses from the laboratory repository of Ontario, Canada's most populous province. Respiratory viruses account for a substantial burden of disease globally [21,22], causing both respiratory and nonrespiratory illnesses [23]. It is impossible to distinguish which respiratory virus is causing infection based on clinical examination alone, necessitating laboratory testing for confirmation. We sought to (1) implement a deep learning–based NLP predictive model to extract respiratory virus information from the laboratory repository and (2) evaluate the generalizability and robustness of predictions (extracted information) across different categories of respiratory viruses and test sets. Our study findings can inform public health practitioners and researchers about using NLP approaches to empower and facilitate access and retrieval of information from a collection of text-based laboratory reports without any time-consuming handcrafted rule-based approaches. This can facilitate the development of a scalable and easily deployable automated information extraction tool.

Figure 1. Semi-rule-based workflow versus fully automated deep learning natural language processing (NLP) approach. Semi-rule-based relies on time-consuming and operationally resource-intensive approaches for the information extraction task. The corpus was derived from the Ontario Laboratories Information System (OLIS). Following basic text-cleaning steps, around 87,500 unique laboratory reports were collected and included in our corpus to be used in parallel by both semi-rule-based and deep learning NLP approaches. Semi-rule-based workflow is a multistep procedure where all the unique reports were grouped by Logical Observation Identifiers Names and Codes (LOINC), year, and location in the first step. In the second step, subject matter experts (SMEs) created a list of dictionaries for terms related to the different viruses and strains and a set of if-then-else rules to generate interpretations and extract information from each laboratory report. The dictionaries and if-then-else rules were packaged as a python library called the rule-based text parser. Finally, the parser was improved based on inputs from 3 SMEs in an iterative manner.



Methods

Study Design

The data set used in this study was a collection of laboratory reports that covered testing for 14 different respiratory viruses and viral subtypes (Table 1), most of which were in the form of texts. The reports were text-based and required cleaning, parsing, and encoding.

The data set was derived from the Ontario Laboratories Information System (OLIS). OLIS has over 100 contributors, which comprise hospital, commercial, and public health laboratories, adding to the complexity and variability of the clinical data. These data were analyzed at ICES.

The automated encoding of laboratory testing reports into respiratory viruses is framed as a multilabel hierarchical classification task to address the needs of knowledge users in

our institute in distinguishing respiratory viruses. According to our users, information at 2 resolution levels is needed: high and low. Therefore, we defined 2 levels of a classification hierarchy, and at each level, the classification was multilabel. Each input text sequence was assigned to a nonempty subset of various labels, as shown in Figure 2. In the first level of the hierarchy, the classifier assigned outputs to 24 mutually nonexclusive fine-grained labels. The fine-grained labels were reassigned to 6 coarse-grained sets of labels in the second level of the classification hierarchy. In this work, “sequence” refers to the input laboratory reports to the NLP model, which may be single or several sentences. A “label” also refers to a status of a specific virus or strain being tested or detected.

To summarize, the information extraction for an input text sequence involved retrieving virus types and identifying their status as being tested and/or detected. Figure 2 illustrates a running example of the input and output of the deep learning-based NLP model.

Table 1. Details of the respiratory viruses embedded in text-based laboratory reports derived from the Ontario Laboratories Information System (OLIS). Specimens may be tested for 1 or more of the following viruses: influenza, RSV^a, adenoviruses, seasonal coronaviruses, enterovirus/rhinoviruses, parainfluenza viruses, HMV^b, and bocavirus^c.

Viruses	Mention counts ^d , n (%)	Tested ^e , n (%)	Detected ^f , n (%)
Adenovirus	21,614 (7)	45 (6)	2 (1)
Bocavirus	5112 (2)	96 (13)	5 (3)
Coronavirus (seasonal)	9128 (3)	95 (13)	9 (5)
Any influenza	49,282 (16)	78 (11)	35 (20)
Influenza A	44,753 (15)	80 (11)	30 (18)
Influenza A H1	6797 (2)	N/A ^g	17 (10)
Influenza A H3	9929 (3)	N/A	18 (10)
Influenza B	40,840 (13)	78 (11)	12 (7)
Enterovirus/rhinovirus	13,262 (4)	92 (13)	19 (11)
HMV	21,194 (7)	46 (6)	3 (2)
Parainfluenza	21,584 (7)	46 (6)	4 (2)
Any RSV	38,080 (12)	68 (9)	11 (6)
RSV A	11,227 (4)	N/A	2 (1)
RSV B	11,094 (4)	N/A	3 (2)
Total	303,896 (100)	724 (100)	170 (100)

^aRSV: respiratory syncytial virus.

^bHMV: human metapneumovirus.

^cThe testing modalities employed include single and multiplex polymerase chain reaction (PCR), direct fluorescent antibody, viral culture, and enzyme immunoassay rapid antigen tests. Repeated testing may involve multiple laboratories and testing modalities.

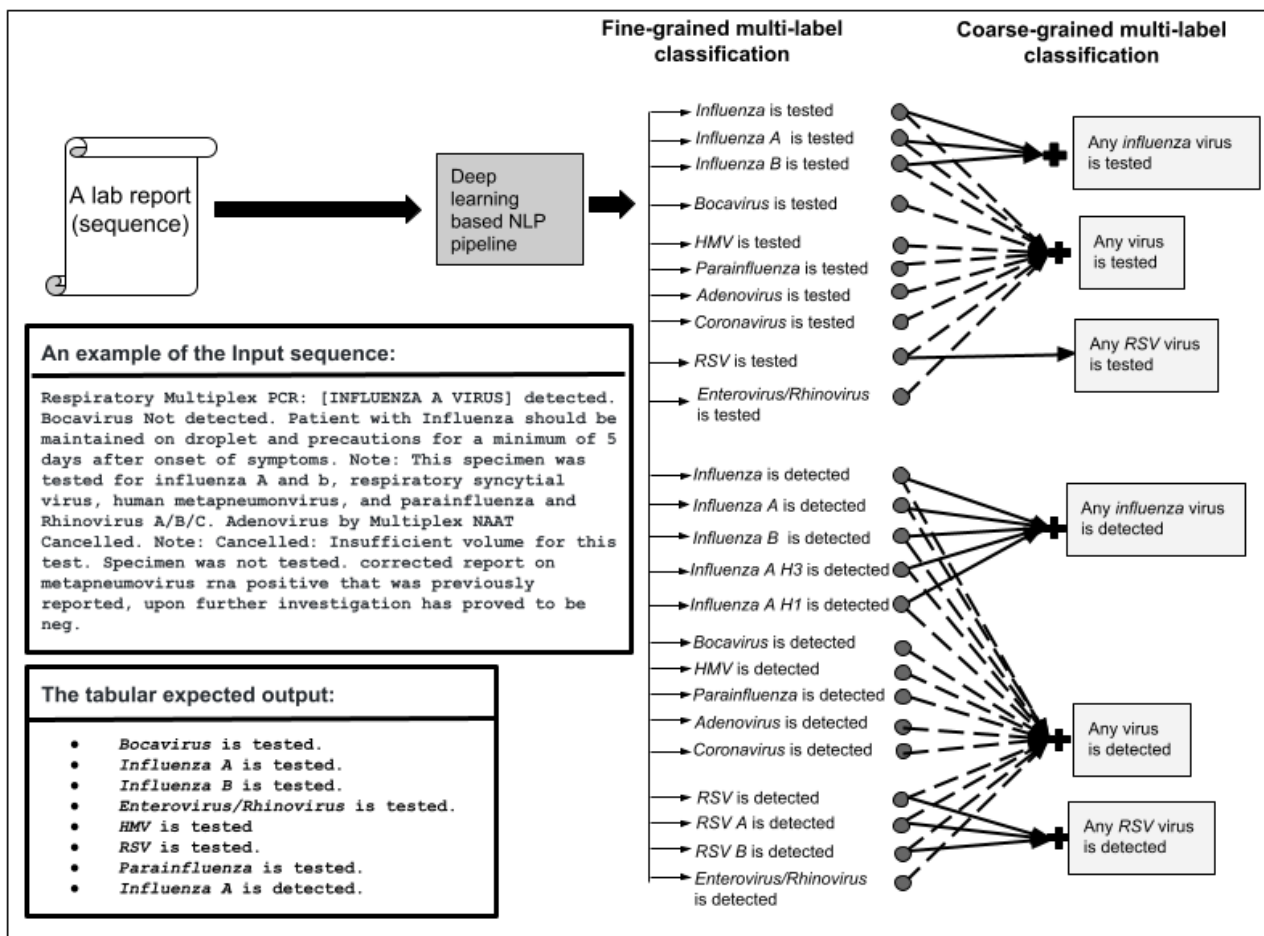
^dRepresents the counts of specific virus terms from all the distinct laboratory reports (unique sequences). It does not provide any clinical information regarding the prevalence of the aforementioned viruses in Ontario.

^eRepresents the proportion of mentions flagged as tested by the parser.

^fRepresents the proportion of mentions flagged as positively detected by the parser. Note that tested and detected are not mutually exclusive; we first determined whether it was tested for (ie, has a result) and then flagged it as detected if the result is positive. Detected is a subset of the tested.

^gN/A: not applicable. Note that the subtypes of influenza A and RSV were only analyzed for detection but not testing, as the scope of the planned analyses for using the respiratory virus data was primarily focused on the larger virus categories.

Figure 2. The fully automated deep learning–based natural language processing (NLP) approach is a hierarchical-based multilabel classification task that retrieves virus (or strain) types and identifies their status as being tested and/or detected. Note that a sequence refers to the input laboratory reports to the NLP approach, which may be a single or several sentences. A label also refers to the status of a specific virus or strain (tested or detected). “influenza is tested” implies it was tested for any influenza type; however, the total number of “influenza is tested” is greater than the total number of “influenza A tested + influenza B tested” since not all influenza types are mentioned. The same applies to “influenza is detected” and “RSV is tested.” HMV: human metapneumovirus; NAAT: nucleic acid amplification test; PCR: polymerase chain reaction; RSV: respiratory syncytial virus.



Corpus Development Description

About OLIS

To create the corpus for this study, over a million observations corresponding to 99 unique Logical Observation Identifiers Names and Codes (LOINC) were pulled from OLIS, and the text-based laboratory results were extracted from the observations. OLIS was created and is managed by Ontario Health, from whom ICES receives an ongoing data feed. At the time of writing this paper, the OLIS data held at ICES consists of >9000 unique LOINC and >5 billion laboratory observations across 150 laboratory test centers in Ontario. As such, the clinical laboratory data have considerable complexity and variability.

Development of the Ground Truth

In this study, we leveraged the semi–rule-based workflow, an information extraction workflow relying on a rule-based and handcrafted tools library, to create ground truth for the deep learning model. A group of 8 SMEs was engaged in performing the required tasks in the workflow; they comprised 2 infectious disease epidemiologists (authors JCK and SAB), 2 infectious disease microbiologists (AM and SM), a genomic specialist

(AMA), a research methodologist (MA), a data analyst (BC), and a machine learning scientist (ED). These tasks included basic text cleaning, quality checking, and rule-based algorithm development for interpreting reports, as shown in Figure 1. In our institute, LOINC are mainly used to filter OLIS observations into relevant groupings (eg, respiratory viruses) and not for encoding and interpretation since they are not always used appropriately by those entering the data into OLIS. Consequently, the SMEs identified a list of 99 LOINC related to respiratory viruses, and all the laboratory reports in OLIS corresponding to these LOINC were retrieved. The workflow consists of 3 tasks, which are detailed in the subsequent paragraphs.

First, the data analyst and data scientist (authors BC and ED) scanned the text strings. After performing basic text cleaning (eg, removing punctuations, stop words, case normalization, lemmatization, and stemming) and removing duplicates, they created a meaningful list of 87,500 unique laboratory reports.

Next, the unique reports were grouped by laboratory and facility names, LOINC, and year. Then, 3 SMEs, including 2 analysts and an infectious disease specialist, manually reviewed multiple samples per group and created a knowledge base and sets of

if-then-else rules to generate interpretations for each laboratory report. Specifically, the knowledge base consisted of dictionaries for terms related to the different viruses and strains. The if-then-else rules provided instructions for grouping virus terms with respective results packaged as a Python library, which we refer to in this study as the rule-based text parser.

Following the initial development of the rule-based text parser, it was improved based on inputs from 3 other SMEs in an iterative manner. The text parser was applied to the entire corpus to generate annotations at each iteration. Next, the data analyst manually reviewed the interpretations and flagged unclear results to be reviewed by SMEs at another iteration. In addition, a small random sample of unflagged test results was provided to SMEs to be reviewed at this iteration. The SMEs subsequently reviewed the list and provided new rules to be added to the text parser. This procedure was repeated until there were no more flagged test results.

Model Development and Evaluation

NLP Model Description

The deep learning-based NLP model consisted of 3 components that were trained jointly: the word embedding layer, the Bi-LSTM layer, and the output layer. The word embedding layer computed a vector representation of each word in the text as a combination of a character-based representation learning model [24,25] and word vectors initialized with pretrained global

vectors (GloVe) embeddings [26]. The embedding layer was coupled with a Bi-LSTM on top of it to generate conceptually and contextually meaningful representations of words. An output layer of a size equal to the number of distinct labels was placed on top of Bi-LSTM, and the last hidden state of the Bi-LSTM was projected into the output layer.

Model Evaluation

The model's robustness and generalizability were evaluated internally and externally on various test sets, as shown in Table 2. The internal test set used for model training was a randomly sampled subset representing 10% (n=6719) of the laboratory reports from OLIS from 2007 to 2018. The performance of the model was also evaluated on 2 out-of-time test sets, including samples from an entirely different time period: (1) a large pre-COVID-19 (2019) sample and (2) a small post-COVID-19 (2020) sample. A separate test set, denoted as the external test set, included samples up to 2019 from 2 separate laboratories (testing sites not included in the development of the model) and was used to assess the external generalizability of the model. F_1 -scores, along with precision and recall scores, were calculated for the model's predictions. A 2-tailed paired t test was used to determine whether there was a statistically significant difference in the F_1 -scores between classes and test sets. In addition, 95% CIs were calculated for the precision and recall scores to quantify the uncertainty of the model's estimates.

Table 2. Data set statistics for laboratory descriptions of the development and test sets.

Cohorts	Sequences ^a , n (%)	Any influenza virus ^b		Any RSV ^c virus		Any virus	
		Detected, n (%)	Tested, n (%)	Detected, n (%)	Tested, n (%)	Detected, n (%)	Tested, n (%)
Total	87411 (100%)						
Development set (2009-2018)							
Training set	60,471 (69)	13,792 (16)	35,292 (40)	3959 (4)	27,196 (31)	22,284 (25)	40,652 (46)
Internal test set	6719 (8)	1604 (2)	3941 (4)	428 (0.5)	3009 (3)	2541 (3)	4534 (5)
Out-of-time test sets							
Pre-COVID-19 (2019)	15,908 (18)	3019 (3)	6903 (8)	706 (0.8)	5957 (7)	4745 (5)	8643 (10)
Post-COVID-19 (2020)	100 (0.01)	N/A ^d	11 (0.01)	<6 (0.006)	11 (0.01)	<6 (0.006)	27 (0.03)
External test set (2009-2018)	4213 (5)	864 (1)	3020 (34)	261 (0.2)	2546 (3)	1431 (2)	3237 (4)

^aRepresents the counts of unique sequences; a sequence refers to the input laboratory reports to the NLP model, which may be a single sentence or several sentences.

^bDetected and tested represent the aggregation of the proportion of any mentions of the virus terms from the total unique sequences in the data set.

^cRSV: respiratory syncytial virus.

^dN/A: not applicable.

Ethical Considerations

The use of the data in this study was approved by the ICES Privacy and Legal Office. Projects that solely use data collected by ICES under section 45 of Ontario's Personal Health Information Protection Act (PHIPA) are exempt from research ethics board review. Section 45 of the PHIPA authorizes ICES to collect personal health information, without consent for the purpose of analyzing or compiling statistical information

concerning the management, evaluation, monitoring, and allocation of resources to or planning for the health system.

Results

The development corpus, including training and test sets, included 87,500 sequences involving ~5 million tokens. The summary statistics for the data sets are shown in Table 2. The NLP model was implemented in TensorFlow on an NVidia

Tesla (Nvidia) graphics processing unit, and Adam optimization was used as the optimization algorithm (more details in [Multimedia Appendix 1](#)). The maximum sequence length was fixed to 400 words. The model was trained several times with random initialization on the development corpus, and the results of the top 10 best-performing models on the test sets are presented in this paper. The results for the fine-grained classification in the first level of the hierarchy are presented in [Table 3](#) and aggregated by microaveraging across the 24 fine-grained labels. Detailed performance for each label is also shown in [Multimedia Appendix 2](#). The F_1 -score performance of the model in the second level of the hierarchy, coarse-grained multilabel classification, for “any influenza,” “any RSV” (respiratory syncytial virus), and “any virus” are shown in [Table 3](#). In addition, the variation of the model’s precision and recall scores using bar plots and 95% CIs are shown in [Figure 3](#).

As expected, the performance on the internal test set was better than the out-of-time (pre-COVID-19) and external test sets. In this regard, the F_1 -score results of the test sets were compared, and noticeable differences were observed between the pairs of internal and out-of-time (pre-COVID-19) test sets, internal and out-of-time (post-COVID-19) test sets, and internal and external test sets. The out-of-time (post-COVID-19) test set was a small

and imbalanced sample, including 100 sequences with <6 mentions of any virus as being detected. The sample included 12 sequences labeled as being tested for coronavirus, and our model correctly classified them with an F_1 -score of 0.67. Regarding the degree of uncertainty in the estimates, fewer variations in precision and recall scores are observed for the internal and out-of-time test sets (pre-COVID-19). On the contrary, the estimates on the out-of-time (post-COVID-19) and external test sets have larger CIs.

In general, the models’ estimates on any test sets were variable across classes with varying degrees of uncertainty. The averaged F_1 -scores of the estimates for both fine-grained (microaveraged) and “coarse-grained any virus” classes were above 90% on the internal test set. The F_1 -score for the “coarse-grained any influenza detected” on all test sets was above 91%. Overall, the performance for coarse-grained detected classes was lower than for coarse-grained tested classes. Among the detected classes, the performance for “any influenza virus” was evidently higher than “any RSV virus.” The same result was observed between “any influenza virus” and “any RSV virus.” Comparably, larger CIs are evidenced for the “coarse-grained any RSV detected” estimates.

Table 3. The prediction results (F_1 -score) of the top 10 best-performing models on the in-time, out-of-time, and external test sets. The fine-grained results are aggregated by microaveraging across 24 fine-grained labels.

Variables	Internal test set	Out-of-time test set ^a		External test set
		(Pre-COVID-19)	(Post-COVID-19)	
Fine-grained microaveraged, mean (SD)	97.3 (0.25)	94.31 (0.59)	60.45 (7.99)	96.23 (0.38)
Coarse-grained any influenza virus, mean (SD)				
Detected ^b	97.64 (0.28)	94.47 (1.04)	N/A ^c	91.11 (2.14)
Tested ^b	98.71 (0.15)	97.26 (0.45)	69.8 (4.43)	98.94 (0.1)
Coarse-grained any RSV^d, mean (SD)				
Detected	90.94 (1.7)	81.56 (3.63)	48.33 (44.76)	57.68 (12.53)
Tested	98.16 (0.34)	96.18 (0.95)	95.6 (5.69)	98.02 (0.47)
Coarse-grained any virus, mean (SD)				
Detected	95.01 (1)	92.31 (1.59)	31.71 (9.44)	82.83 (3.27)
Tested	98.4 (0.17)	96.3 (0.35)	75.87 (4.82)	98.59 (0.2)

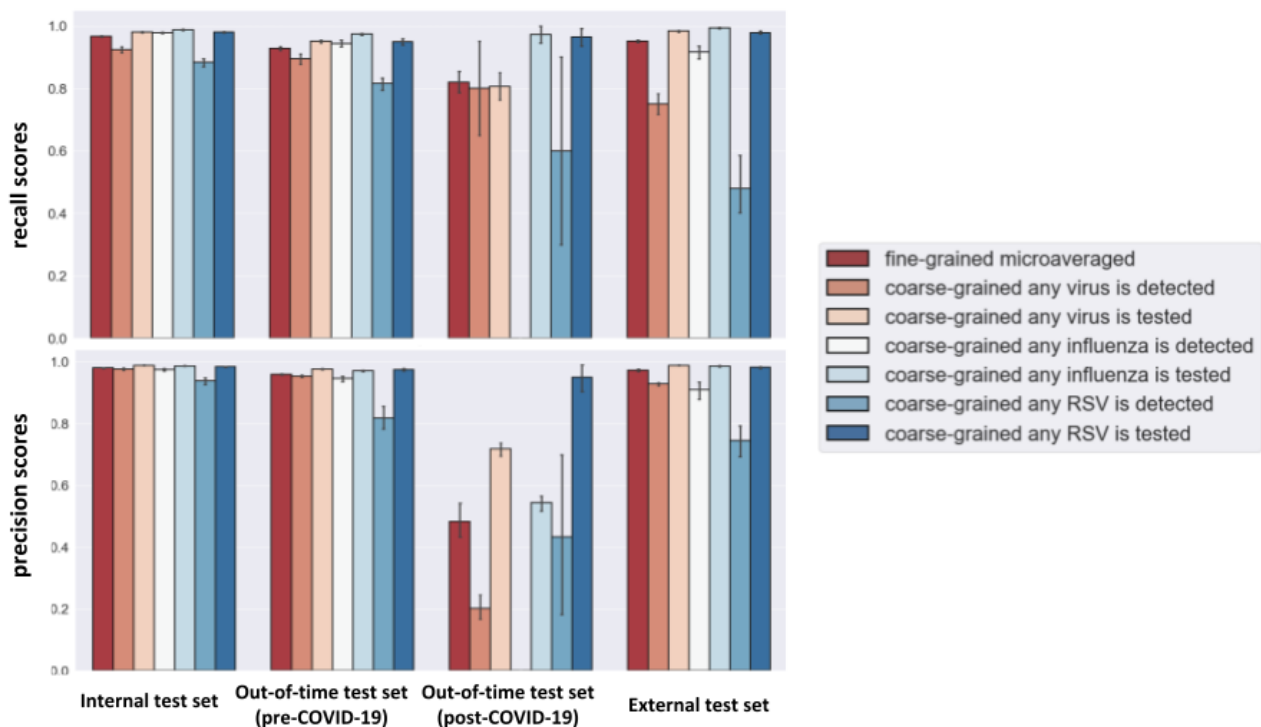
^aThe out-of-time test set (post-COVID-19) is a very small and imbalanced sample, including only 100 sequences with no mentions of any virus detected.

^bDetected and tested represent the aggregation of the proportion of any mentions of the virus terms from the total unique sequences in the data set.

^cN/A: not applicable.

^dRSV: respiratory syncytial virus.

Figure 3. The precision and recall scores of the predictions of the top 10 best-performing models with 95% CIs. The fine-grained results are aggregated by microaveraging across 24 fine-grained labels. RSV: respiratory syncytial virus.



Discussion

Principal Findings

In this study, we demonstrated an implementation and evaluation of an NLP model for an automated and reductive information extraction task in a province-wide laboratory data repository. Our results suggest that the NLP model is a promising approach for information extraction from text-based laboratory reports as an alternative method to address the time-consuming and operationally resource-intensive nature of handcrafted rule-based models.

Overview of Model Findings

Generalization Across Various Test Sets

Overall, the NLP solution, which was a hierarchical multilabel classifier, performed well on the internal, out-of-time (pre-COVID-19), and external (different laboratories) test sets. Except for the internal test sets, the other test sets were sourced from either a more recent time period or other laboratory sites, but the model was able to generalize well with microaveraged F_1 -score >94% across all classes. The performance of the model on the other out-of-time (post-COVID-19) test set was satisfactory; however, due to its small sample size with many underrepresented classes, it was not possible to draw any conclusion. The out-of-time (post-COVID-19) test set was pulled from the 2020 cohort to simulate a nonstationary production environment for observation.

Stability and Performance Variation Between Classes

In general, the model's performance on any test sets was variable across classes and virus types due to the imbalanced nature of the corpus and sample sizes per class. There were intrinsically

fewer classes of viruses detected compared with those tested. Therefore, the model's performance was noticeably lower in the "detected" cases. Among the detected cases, the lowest performance was observed for RSV, and the highest performance among the tested cases was observed for influenza. Moreover, more considerable variations were observed for the positive predictive and sensitivity values of the detected classes, particularly for the "any RSV virus detected" class.

Comparison With Prior Work

Deep learning-based NLP approaches have demonstrated efficacy in many clinical NLP tasks and have thoroughly permeated the informatics community. The existing body of literature has mainly focused on using deep learning models to extract and interpret cancer-related clinical concepts [17,27,28] from free text or other clinically meaningful entities from radiology reports or hospital notes [10,15]. At the time of writing this paper, only 1 study has explored the use of an NLP system, Topaz, for the automated extraction and classification of influenza-related terms from text emergency reports [29-31]. To our knowledge, our study is the first to explore using deep learning models for efficient processing and extraction of clinically meaningful knowledge pertaining to respiratory viruses from a laboratory repository.

One strength of the NLP approach used in this study is its scalability for various text-based laboratory scenarios. As the size and complexity of laboratory data grow, so does the need for scalable and reusable tools for automated extraction of knowledge from vast amounts of clinical notes and quick generalization from 1 task to another. Manual processing of laboratory reports severely limits the utilization of rich information embedded in the data repositories and makes the process of data cleaning and quality improvement prohibitively

expensive. Usually, the rules learned from cleaning a single collection of laboratory reports show little generalizability toward other collections. On the other hand, deep learning-based NLP algorithms are well poised to scale the information extraction process. Although building deep learning-based NLP models is computationally intensive and memory demanding, the benefit-to-cost ratio of these models in clinical settings will continue to increase.

Limitations

Although this deep learning model promises great potential for digitized health data, putting the model into production and prospectively validating operational data is as crucial as model building and a critical step in assessing and ensuring its operational effectiveness. However, we expect the model's performance to deteriorate as it goes into production, potentially impacting data quality. Moving forward, we plan to run a silent-period production validation to further prospectively explore the model's performance. During the silent period, our model will be integrated into the data quality and management workflow for the laboratory data repository, and the outputs will be internally validated in a fashion that would avoid exposure to data users. We also plan to run rigorous evaluation and continuous refinement of the model in the silent period to assess its performance better before it enters production. Transformers heralded a new era in the NLP field and have shown to be very successful in many tasks. Our future direction includes improving the performance of our NLP pipeline by adding transformer models.

Another significant limitation of this study is that the model was only trained on respiratory virus laboratory reports. Even within that collection, some categories were naturally underrepresented, which impacted the model's generalizability. Therefore, during the silent period, more records from a diverse set of laboratory reports from various categories will be annotated and made available to the model, and the model will be updated accordingly. Finally, this study lacks explainability, which could limit the adoption of our deep learning-based models in future applications. Therefore, we plan to develop parallel pipelines that help explain the representations of the laboratory reports and the classifier's decision boundary.

Conclusion

The health industry is rapidly becoming digitized, and information extraction is a promising method for researchers and clinicians seeking quick retrieval of information embedded in texts. This study described developing and validating a deep learning-based NLP approach to extract respiratory virus testing information from laboratory reports. We demonstrated that our system could classify and encode large volumes of text-based laboratory reports with high performance without any of the previous time-consuming handcrafted feature engineering approaches. Taken together, the findings of this study provide encouraging support that NLP-based information extraction could become an important component of laboratory information repositories to assist researchers, clinicians, and health care providers with their information and knowledge management tasks.

Acknowledgments

This study was a collaborative effort supported by the Vector Institute, an independent, not-for-profit corporation dedicated to research in the field of artificial intelligence, and ICES, an independent, nonprofit research organization that uses population-based health and social data to produce knowledge on a broad range of health care issues. Resources used in preparing this work were funded by an annual grant from the Ontario Ministry of Health (MOH) and the Ministry of Long-Term Care (MLTC). Parts of this material are based on data and information compiled and provided by the Ontario Ministry of Health. This work was also supported by a SickKids-Canadian Institutes of Health Research New Investigator Grant in Child and Youth Health (NI19-1065). The analyses, conclusions, opinions, and statements expressed herein are solely those of the authors and do not reflect those of the funding or data sources; no endorsement is intended or should be inferred.

Data Availability

The data underlying this work are held securely in coded form at ICES and therefore cannot be shared publicly due to data privacy concerns and legal data sharing agreements between ICES and data providers (eg, health care organizations and the government). However, data access might be granted to those who meet prespecified criteria for confidential access (email: das@ices.on.ca).

Conflicts of Interest

None declared.

Multimedia Appendix 1

Details of hyperparameter tuning.

[[PDF File \(Adobe PDF File\), 51 KB - ai_v2i1e44835_app1.pdf](#)]

Multimedia Appendix 2

Fine-grained classification results (F1-scores from the best performing model).

[[PDF File \(Adobe PDF File\), 83 KB - ai_v2i1e44835_app2.pdf](#)]

References

1. Abhyankar S, Demner-Fushman D, McDonald CJ. Standardizing clinical laboratory data for secondary use. *J Biomed Inform* 2012 Aug;45(4):642-650 [FREE Full text] [doi: [10.1016/j.jbi.2012.04.012](https://doi.org/10.1016/j.jbi.2012.04.012)] [Medline: [22561944](https://pubmed.ncbi.nlm.nih.gov/22561944/)]
2. Kudler NR, Pantanowitz L. Overview of laboratory data tools available in a single electronic medical record. *J Pathol Inform* 2010 May 26;1(1):3 [FREE Full text] [doi: [10.4103/2153-3539.63824](https://doi.org/10.4103/2153-3539.63824)] [Medline: [20805960](https://pubmed.ncbi.nlm.nih.gov/20805960/)]
3. Clarke W, Marzinke M, editors. *Contemporary Practice in Clinical Chemistry*. Cambridge, MA: Academic Press; Jun 08, 2020.
4. Ross MK, Wei W, Ohno-Machado L. "Big Data" and the electronic health record. *Yearb Med Inform* 2018 Mar 05;23(01):97-104. [doi: [10.15265/iy-2014-0003](https://doi.org/10.15265/iy-2014-0003)]
5. Jensen PB, Jensen LJ, Brunak S. Mining electronic health records: towards better research applications and clinical care. *Nat Rev Genet* 2012 May 02;13(6):395-405. [doi: [10.1038/nrg3208](https://doi.org/10.1038/nrg3208)] [Medline: [22549152](https://pubmed.ncbi.nlm.nih.gov/22549152/)]
6. Kuo TT, Rao P, Maehara C, Doan S, Chaparro JD, Day ME, et al. Ensembles of NLP tools for data element extraction from clinical notes. 2017 Presented at: AMIA Annual Symposium; November 16; Chicago, IL p. 1880-1889. [doi: [10.5281/zenodo.1491953](https://doi.org/10.5281/zenodo.1491953)]
7. 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 Sep 01;17(5):507-513 [FREE Full text] [doi: [10.1136/jamia.2009.001560](https://doi.org/10.1136/jamia.2009.001560)] [Medline: [20819853](https://pubmed.ncbi.nlm.nih.gov/20819853/)]
8. Friedman C, Shagina L, Lussier Y, Hripcsak G. Automated encoding of clinical documents based on natural language processing. *J Am Med Inform Assoc* 2004 Sep 01;11(5):392-402. [doi: [10.1197/jamia.m1552](https://doi.org/10.1197/jamia.m1552)]
9. Huang K, Altsaar J, Ranganath R. ClinicalBERT: modeling clinical notes and predicting hospital readmission. arXiv Preprint posted online April 10, 2019. [FREE Full text] [doi: [10.48550/arXiv.1904.05342](https://doi.org/10.48550/arXiv.1904.05342)]
10. Kormilitzin A, Vaci N, Liu Q, Nevado-Holgado A. Med7: A transferable clinical natural language processing model for electronic health records. *Artif Intell Med* 2021 Aug;118:102086. [doi: [10.1016/j.artmed.2021.102086](https://doi.org/10.1016/j.artmed.2021.102086)] [Medline: [34412834](https://pubmed.ncbi.nlm.nih.gov/34412834/)]
11. Wu Y, Jiang M, Xu J, Zhi D, Xu H. Clinical named entity recognition using deep learning models. 2017 Presented at: AMIA Annual Symposium; November 16; Washington, DC p. 1812-1819.
12. Si Y, Wang J, Xu H, Roberts K. Enhancing clinical concept extraction with contextual embeddings. *J Am Med Inform Assoc* 2019 Nov 01;26(11):1297-1304 [FREE Full text] [doi: [10.1093/jamia/ocz096](https://doi.org/10.1093/jamia/ocz096)] [Medline: [31265066](https://pubmed.ncbi.nlm.nih.gov/31265066/)]
13. Esteva A, Robicquet A, Ramsundar B, Kuleshov V, DePristo M, Chou K, et al. A guide to deep learning in healthcare. *Nat Med* 2019 Jan 7;25(1):24-29. [doi: [10.1038/s41591-018-0316-z](https://doi.org/10.1038/s41591-018-0316-z)] [Medline: [30617335](https://pubmed.ncbi.nlm.nih.gov/30617335/)]
14. Wu S, Roberts K, Datta S, Du J, Ji Z, Si Y, et al. Deep learning in clinical natural language processing: a methodical review. *J Am Med Inform Assoc* 2020 Mar 01;27(3):457-470 [FREE Full text] [doi: [10.1093/jamia/ocz200](https://doi.org/10.1093/jamia/ocz200)] [Medline: [31794016](https://pubmed.ncbi.nlm.nih.gov/31794016/)]
15. Boag W, Wacome K, Naumann T, Rumshisky A. CliNER: A lightweight tool for clinical named entity recognition. 2015 Presented at: AMIA Joint Summits on Clinical Research Informatics; March 23-27; San Francisco, CA.
16. Sugimoto K, Takeda T, Oh J, Wada S, Konishi S, Yamahata A, et al. Extracting clinical terms from radiology reports with deep learning. *J Biomed Inform* 2021 Apr;116:103729 [FREE Full text] [doi: [10.1016/j.jbi.2021.103729](https://doi.org/10.1016/j.jbi.2021.103729)] [Medline: [33711545](https://pubmed.ncbi.nlm.nih.gov/33711545/)]
17. Si Y, Roberts K. A frame-based NLP system for cancer-related information extraction. 2018 Presented at: AMIA Annual Symposium; November 3-7; San Francisco, CA URL: <https://pubmed.ncbi.nlm.nih.gov/30815198/>
18. Banerjee I, Ling Y, Chen MC, Hasan SA, Langlotz CP, Moradzadeh N, et al. Comparative effectiveness of convolutional neural network (CNN) and recurrent neural network (RNN) architectures for radiology text report classification. *Artif Intell Med* 2019 Jun;97:79-88 [FREE Full text] [doi: [10.1016/j.artmed.2018.11.004](https://doi.org/10.1016/j.artmed.2018.11.004)] [Medline: [30477892](https://pubmed.ncbi.nlm.nih.gov/30477892/)]
19. Chen MC, Ball RL, Yang L, Moradzadeh N, Chapman BE, Larson DB, et al. Deep learning to classify radiology free-text reports. *Radiology* 2018 Mar;286(3):845-852. [doi: [10.1148/radiol.2017171115](https://doi.org/10.1148/radiol.2017171115)] [Medline: [29135365](https://pubmed.ncbi.nlm.nih.gov/29135365/)]
20. Gao S, Young MT, Qiu JX, Yoon HJ, Christian JB, Fearn PA, et al. Hierarchical attention networks for information extraction from cancer pathology reports. *J Am Med Inform Assoc* 2018 Mar 01;25(3):321-330 [FREE Full text] [doi: [10.1093/jamia/ocz131](https://doi.org/10.1093/jamia/ocz131)] [Medline: [29155996](https://pubmed.ncbi.nlm.nih.gov/29155996/)]
21. GBD 2013 Mortality Causes of Death Collaborators. Global, regional, and national age-sex specific all-cause and cause-specific mortality for 240 causes of death, 1990-2013: a systematic analysis for the Global Burden of Disease Study 2013. *Lancet* 2015 Jan 10;385(9963):117-171 [FREE Full text] [doi: [10.1016/S0140-6736\(14\)61682-2](https://doi.org/10.1016/S0140-6736(14)61682-2)] [Medline: [25530442](https://pubmed.ncbi.nlm.nih.gov/25530442/)]
22. GBD 2016 Disease Injury Incidence Prevalence Collaborators. Global, regional, and national incidence, prevalence, and years lived with disability for 328 diseases and injuries for 195 countries, 1990-2016: a systematic analysis for the Global Burden of Disease Study 2016. *Lancet* 2017 Sep 16;390(10100):1211-1259 [FREE Full text] [doi: [10.1016/S0140-6736\(17\)32154-2](https://doi.org/10.1016/S0140-6736(17)32154-2)] [Medline: [28919117](https://pubmed.ncbi.nlm.nih.gov/28919117/)]
23. Macias AE, McElhaney JE, Chaves SS, Nealon J, Nunes MC, Samson SI, et al. The disease burden of influenza beyond respiratory illness. *Vaccine* 2021 Mar 15;39 Suppl 1:A6-A14 [FREE Full text] [doi: [10.1016/j.vaccine.2020.09.048](https://doi.org/10.1016/j.vaccine.2020.09.048)] [Medline: [33041103](https://pubmed.ncbi.nlm.nih.gov/33041103/)]
24. Lample G, Ballesteros M, Subramanian S, Kawakami K, Dyer C. Neural architectures for named entity recognition. : Association for Computational Linguistics; 2016 Presented at: Conference of the North American Chapter of the Association

- for Computational Linguistics: Human Language Technologies; June 12-17; San Diego, CA URL: <https://aclanthology.org/N16-1030> [doi: [10.18653/v1/N16-1030](https://doi.org/10.18653/v1/N16-1030)]
25. Kim Y, Jernite Y, Sontag D, Rush A. Character-aware neural language models. Association for the Advancement of Artificial Intelligence 2016 Mar 05;30(1). [doi: [10.1609/aaai.v30i1.10362](https://doi.org/10.1609/aaai.v30i1.10362)]
 26. Pennington J, Socher R, Manning C. Glove: Global vectors for word representation. 2014 Presented at: Conference on Empirical Methods in Natural Language Processing (EMNLP); October 25–29; Doha, Qatar. [doi: [10.3115/v1/D14-1162](https://doi.org/10.3115/v1/D14-1162)]
 27. Zhang X, Zhang Y, Zhang Q, Ren Y, Qiu T, Ma J, et al. Extracting comprehensive clinical information for breast cancer using deep learning methods. Int J Med Inform 2019 Dec;132:103985. [doi: [10.1016/j.ijmedinf.2019.103985](https://doi.org/10.1016/j.ijmedinf.2019.103985)] [Medline: [31627032](https://pubmed.ncbi.nlm.nih.gov/31627032/)]
 28. Savova G, Danciu I, Alamudun F, Miller T, Lin C, Bitterman DS, et al. Use of natural language processing to extract clinical cancer phenotypes from electronic medical records. Cancer Res 2019 Nov 01;79(21):5463-5470 [FREE Full text] [doi: [10.1158/0008-5472.CAN-19-0579](https://doi.org/10.1158/0008-5472.CAN-19-0579)] [Medline: [31395609](https://pubmed.ncbi.nlm.nih.gov/31395609/)]
 29. López Pineda A, Ye Y, Visweswaran S, Cooper GF, Wagner MM, Tsui F. Comparison of machine learning classifiers for influenza detection from emergency department free-text reports. J Biomed Inform 2015 Dec;58:60-69 [FREE Full text] [doi: [10.1016/j.jbi.2015.08.019](https://doi.org/10.1016/j.jbi.2015.08.019)] [Medline: [26385375](https://pubmed.ncbi.nlm.nih.gov/26385375/)]
 30. Ye Y, Tsui F, Wagner M, Espino JU, Li Q. Influenza detection from emergency department reports using natural language processing and Bayesian network classifiers. J Am Med Inform Assoc 2014 Sep 01;21(5):815-823 [FREE Full text] [doi: [10.1136/amiajnl-2013-001934](https://doi.org/10.1136/amiajnl-2013-001934)] [Medline: [24406261](https://pubmed.ncbi.nlm.nih.gov/24406261/)]
 31. Pineda A, Tsui FC, Visweswaran S, Cooper FG. Detection of patients with influenza syndrome using machine-learning models learned from emergency department reports. Online J Public Health Inform 2013 Apr 4;5(1):41-41 [FREE Full text] [doi: [10.5210/ojphi.v5i1.4446](https://doi.org/10.5210/ojphi.v5i1.4446)]

Abbreviations

- Bi-LSTM:** bidirectional long short-term memory
GloVe: global vectors
LOINC: Logical Observation Identifiers Names and Codes
NLP: natural language processing
OLIS: Ontario Laboratories Information System
PHIPA: Personal Health Information Protection Act
RNN: recurrent neural network
RSV: respiratory syncytial virus
SME: subject matter expert

Edited by B Malin, K El Emam, Y Huo; submitted 05.12.22; peer-reviewed by PP Zhao, A Teles; comments to author 10.03.23; revised version received 31.03.23; accepted 18.04.23; published 06.06.23.

Please cite as:

Dolatabadi E, Chen B, Buchan SA, Austin AM, Azimae M, McGeer A, Mubareka S, Kwong JC

Natural Language Processing for Clinical Laboratory Data Repository Systems: Implementation and Evaluation for Respiratory Viruses

JMIR AI 2023;2:e44835

URL: <https://ai.jmir.org/2023/1/e44835>

doi: [10.2196/44835](https://doi.org/10.2196/44835)

PMID: [38875570](https://pubmed.ncbi.nlm.nih.gov/38875570/)

©Elham Dolatabadi, Branson Chen, Sarah A Buchan, Alex Marchand Austin, Mahmoud Azimae, Allison McGeer, Samira Mubareka, Jeffrey C Kwong. Originally published in JMIR AI (<https://ai.jmir.org>), 06.06.2023. This is an open-access article distributed under the terms of the Creative Commons Attribution License (<https://creativecommons.org/licenses/by/4.0/>), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR AI, is properly cited. The complete bibliographic information, a link to the original publication on <https://www.ai.jmir.org/>, as well as this copyright and license information must be included.

Original Paper

Developing an Inpatient Electronic Medical Record Phenotype for Hospital-Acquired Pressure Injuries: Case Study Using Natural Language Processing Models

Elvira Nurmambetova^{1*}, BHSc; Jie Pan^{1,2*}, PhD; Zilong Zhang^{1*}, MSc; Guosong Wu^{1,2}, PhD; Seungwon Lee^{1,2,3}, MPH, PhD; Danielle A Southern^{1,2}, MSc; Elliot A Martin^{1,3}, PhD; Chester Ho⁴, MD; Yuan Xu^{1,2,5,6}, MD, PhD; Cathy A Eastwood^{1,2}, RN, PhD

¹Centre for Health Informatics, Cumming School of Medicine, University of Calgary, Calgary, AB, Canada

²Department of Community Health Sciences, Cumming School of Medicine, University of Calgary, Calgary, AB, Canada

³Alberta Health Services, Edmonton, AB, Canada

⁴Department of Medicine, Faculty of Medicine & Dentistry, University of Alberta, Edmonton, AB, Canada

⁵Department of Oncology, University of Calgary, Tom Baker Cancer Centre, Calgary, AB, Canada

⁶Department of Surgery, Foothills Medical Centre, University of Calgary, Calgary, AB, Canada

*these authors contributed equally

Corresponding Author:

Elvira Nurmambetova, BHSc
Centre for Health Informatics
Cumming School of Medicine
University of Calgary
3330 Hospital Dr NW
Calgary, AB, T2N 4N1
Canada
Phone: 1 4032206843
Email: elvira.nurmambetova@ucalgary.ca

Abstract

Background: Surveillance of hospital-acquired pressure injuries (HAPI) is often suboptimal when relying on administrative health data, as International Classification of Diseases (ICD) codes are known to have long delays and are undercoded. We leveraged natural language processing (NLP) applications on free-text notes, particularly the inpatient nursing notes, from electronic medical records (EMRs), to more accurately and timely identify HAPIs.

Objective: This study aimed to show that EMR-based phenotyping algorithms are more fitted to detect HAPIs than ICD-10-CA algorithms alone, while the clinical logs are recorded with higher accuracy via NLP using nursing notes.

Methods: Patients with HAPIs were identified from head-to-toe skin assessments in a local tertiary acute care hospital during a clinical trial that took place from 2015 to 2018 in Calgary, Alberta, Canada. Clinical notes documented during the trial were extracted from the EMR database after the linkage with the discharge abstract database. Different combinations of several types of clinical notes were processed by sequential forward selection during the model development. Text classification algorithms for HAPI detection were developed using random forest (RF), extreme gradient boosting (XGBoost), and deep learning models. The classification threshold was tuned to enable the model to achieve similar specificity to an ICD-based phenotyping study. Each model's performance was assessed, and comparisons were made between the metrics, including sensitivity, positive predictive value, negative predictive value, and F_1 -score.

Results: Data from 280 eligible patients were used in this study, among whom 97 patients had HAPIs during the trial. RF was the optimal performing model with a sensitivity of 0.464 (95% CI 0.365-0.563), specificity of 0.984 (95% CI 0.965-1.000), and F_1 -score of 0.612 (95% CI of 0.473-0.751). The machine learning (ML) model reached higher sensitivity without sacrificing much specificity compared to the previously reported performance of ICD-based algorithms.

Conclusions: The EMR-based NLP phenotyping algorithms demonstrated improved performance in HAPI case detection over ICD-10-CA codes alone. Daily generated nursing notes in EMRs are a valuable data resource for ML models to accurately detect adverse events. The study contributes to enhancing automated health care quality and safety surveillance.

KEYWORDS

pressure injury; natural language processing; NLP; algorithm; phenotype algorithm; phenotyping algorithm; machine learning; electronic medical record; EMR; pressure sore; pressure wound; pressure ulcer; pressure injuries; detect

Introduction

Pressure injury (PI), also known as a pressure ulcer, is an injury of the skin and deep tissues caused by external pressures. Annually, PIs affect approximately 250,000 to 500,000 Canadians, with an estimated prevalence of 26.0% in health care institutions [1,2]. Hospital-acquired pressure injuries (HAPIs) are PIs developed during an inpatient hospital stay. HAPIs can significantly extend a patient's hospitalization length of stay and cause severe secondary complications, such as muscle and profound tissue impairment [3]. HAPI is considered mostly preventable, and its prevalence has been reckoned as an acceptable indicator of the quality of care [4,5]. Collecting HAPI status using chart review is time and labor-intensive, thereby not suitable for large-scale population-based applications. Considering all the factors, there is a need for automated ways to accurately and timely identify HAPIs for analyzing large cohort studies that support quality improvement efforts and assisting unit managers with developing reliable patient safety programs. The International Classification of Diseases, 10th Revision, adapted to the Canadian health system (ICD-10-CA), can be used to estimate the prevalence of adverse events from administrative data. However, the coded administrative data are prone to miss positive cases: previous research demonstrated that the sensitivity of the ICD algorithm for identifying HAPI cases is around 30% compared to chart review [1]. In addition to the sensitivity issue, ICD codes are not generally assigned with a specific time when diseases occur. Therefore, they are unsuitable for reporting the time when HAPIs occur [6]. Thus, there is a need for more accurate HAPI detection.

Electronic medical records (EMRs) are used to track and organize patient information for efficient treatment of medical conditions in a secure system [7]. Free-text clinical notes in EMRs consist of detailed descriptions of patients' conditions and treatment. Additionally, clinical notes are typically written in a continuous manner across patients' interactions with health care systems, making clinical notes more real-time compared to diagnosis codes. Despite the rich information the clinical record may have, coders often cannot read every entry, given their limited time per chart and many patients have prolonged hospital stays. Recent studies suggest that using free-text in EMRs alone, or incorporating EMR data elements, can significantly improve the accuracy of case identification of specific comorbidities [8-16]. Xu et al compared the ICD algorithm with algorithms based on EMR keyword search, which achieved a high sensitivity of 0.655 (95% CI 0.601-0.710) [8]. The Canadian health system operates as a publicly funded single-payer insurance system by the federal, provincial, and territorial governments [17]. Additional crown institutions at the provincial and federal-level monitor adverse events such as HAPI. For example, in Alberta, Canadian Institute for Health Information, the federal crown corporation, works with Alberta

Health Services (provincial health care agency) to monitor PIs [18,19]. To date, there is no mandatory collection of PIs within Canadian acute-care facilities. Real-time PI evaluation and auditing using ICD codes are not possible as Canadian health data systems are set up such that ICD codes are assigned outside of providing care and have a few months lag in data extraction, transfer, and load [20]. Consequently, these agencies aim to monitor but are unable to conduct real-time auditing of PIs in Canada. Therefore, there is a need to develop EMR data-specific algorithm for identifying PIs for monitoring and auditing within Canadian acute-care facilities. Our objective was to create EMR data-specific algorithms for HAPIs. Availability and implementation of PI-specific algorithms within a clinical information system would allow the abovementioned federal and provincial agencies to conduct real-time surveillance of HAPIs, improving patient safety, enhancing the quality of care, and reducing the burden of costs associated with adverse events. The EMR phenotype case detection is evaluated via comparison with confirmed HAPIs status acquired in a clinical trial [21].

Methods

Study Design

This is an EMR phenotyping study for enhancing HAPI identification using free-text notes. Obtained clinical trial data were linked to administrative and EMR data for model development and validation. The natural language processing (NLP) method's performance was compared with results from the ICD validation study conducted in Alberta, Canada, by Wong et al [21]. Detailed information for HAPIs identification can be found in their study.

Clinical Trial Data

Previously completed randomized controlled trial (RCT) data of 678 eligible consenting inpatients were obtained from an affiliated research team and were used as the reference standard [21]. The trial evaluated the efficacy of a pressure-sensing mattress in preventing interface pressure. A research nurse performed a clinical head-to-toe skin assessment for PI formation, and suspected deep tissue injuries were monitored throughout 3 days of enrollment [21]. Assessments were conducted within 24 hours of admission, on the day of trial enrollment, and the third day after enrollment, and documented in Allscripts Sunrise Clinical Manager (SCM) EMR (Figure 1). Three days were chosen as a length of time for the research nurse to perform data collection and risk assessment for 3 reasons. First, this is the average length of stay in the local inpatient units, and a longer trial period may include varying nursing practice due to hospital discharge with a shorter length of stay, unit changes, and more nursing shift changes. Second, the dedicated investigation team deemed 3 days sufficient for pressure-related skin and soft tissue changes to develop. Lastly, as a continuous collection of interface pressure throughout the

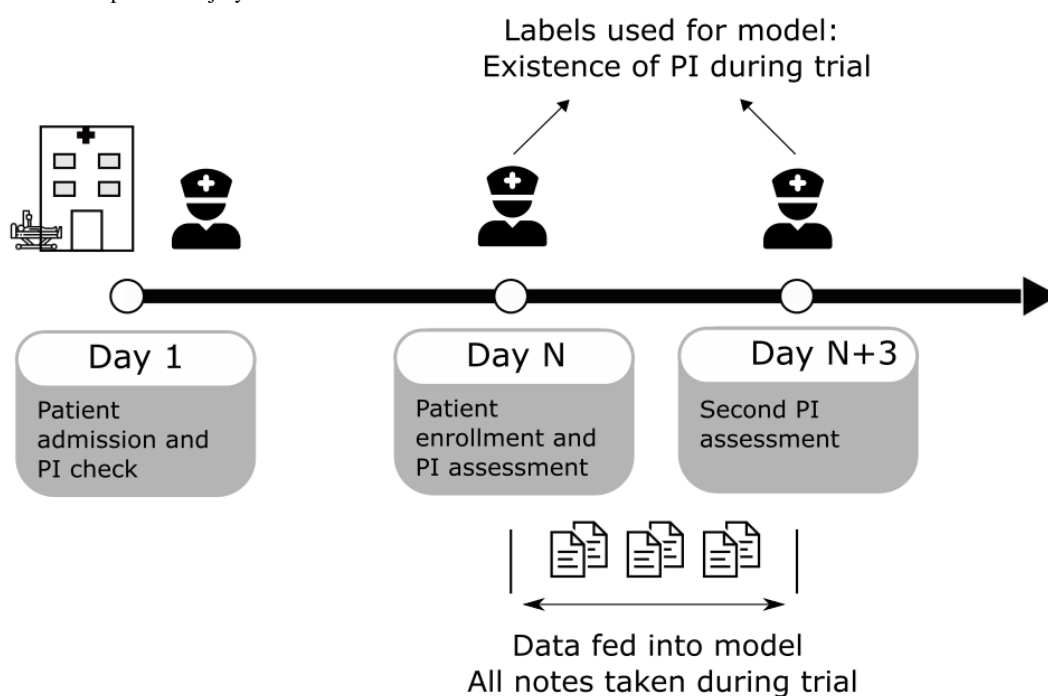
enrollment period leads to a large volume of data, 3 days allowed for optimal data collection while maintaining participant enrollment.

The research nurse, who measured pressure-related skin ulcerations, was trained as a wound care specialist in the provision of pressure ulcers, ostomy, and continence care [21]. The patients' PI status check on admission was determined based on when the patient was admitted. The clinical trial team relied on the medical record if the patient had been admitted long before the study and consented to the study. If the patient agreed to participate in the trial right after being admitted to the hospital, the research nurse noted the PI status on admission.

The following data elements were abstracted from the clinical trial data: record ID, medical unit, sex, first-skin assessment

date, second-skin assessment date, presence of PIs, and other possible related conditions (cerebrovascular disease, diabetes mellitus, etc). The clinical trial measured and classified PIs into 6 stages: stage 1, stage 2, stage 3, stage 4, suspected deep tissue injury, and unstageable PIs [22]. Stages of PIs were identified according to the National Pressure Ulcer Advisory Panel's pressure ulcer staging system [23]. Stage 1 PIs include sores. Stage 2 captures open wounds on the surface of the skin. Stage 3 PIs represent wounds extending beneath the skin and affecting fat tissue. At stage 4, PIs are deep and reach into muscles, bones, and tendons. The trial is registered at clinicaltrials.gov (NCT02325388). Additional details surrounding the clinical trial data were published by Wong et al [21].

Figure 1. Illustration of the clinical trial for assessment of PI status in the enrolled patient cohort (n=678) and the data input used for the development of classification models. PI: pressure injury.



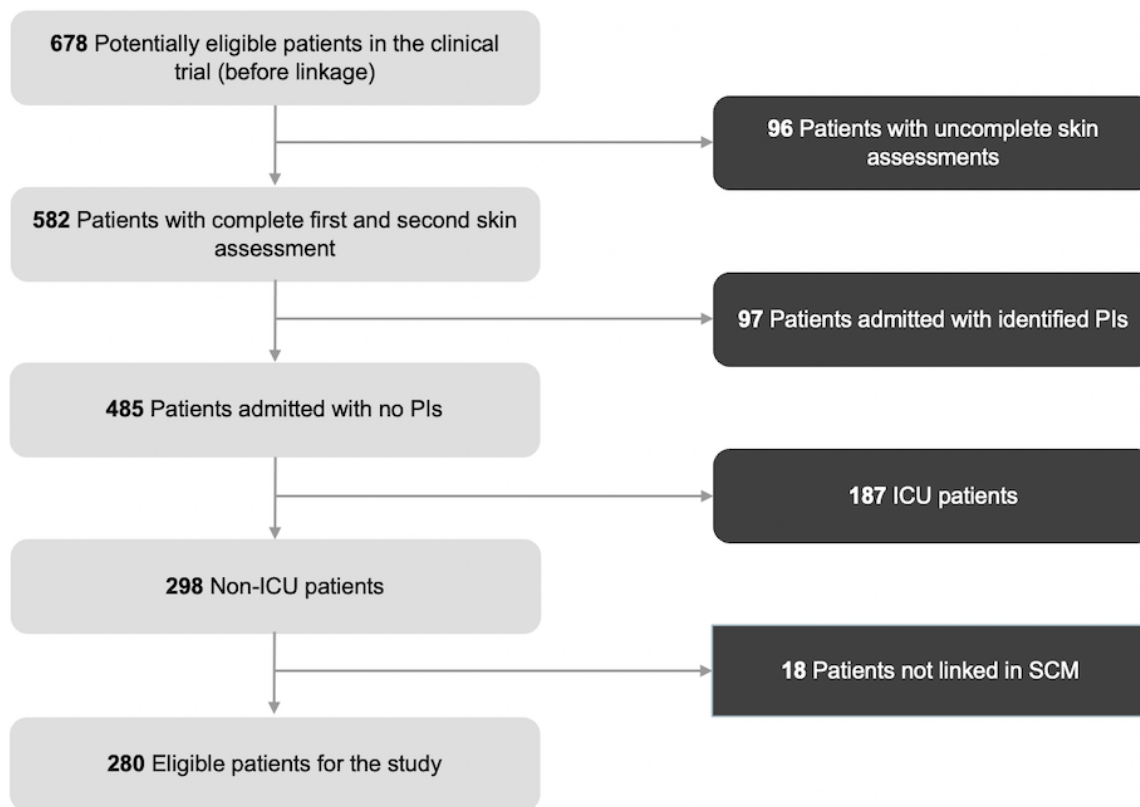
Study Cohort

Inclusion and Exclusion Criteria

During the RCT, eligible patients were at least 18 years old, were expected to have a length of stay of at least 3 days, and did not receive near-end-of-life care within 3 days of trial enrollment [21]. Participants were recruited from nursing units with a high risk for PI development including acute medical, neurosurgery, neurology, and intensive care [21]. For this study,

patients were excluded if their data did not link to EMR data, had incomplete skin assessments, or included erroneous assessment or discharge dates. Patients with PIs on the day of admission were also excluded in order to track only PIs developed during hospitalization. Furthermore, intensive care unit (ICU) patients were excluded since their data were stored in another data warehouse with distinct data elements from those found in SCM and required restricted access. After careful selection, the final cohort of eligible patients was 280 (Figure 2).

Figure 2. Flowchart of inclusion and exclusion criteria for the patient cohort based on trial completeness, PI status, and age (minimum age of 18 years old) for all controls in the panel. ICU: intensive care unit; PI: pressure injury; SCM: Sunrise Clinical Management.



Data Linkage to Discharge Abstract Database and SCM EMR

Deterministic data linkage was performed between the RCT data, administrative data from the discharge abstract database (DAD), and SCM EMR data [24]. SCM was the EMR system employed in Calgary hospitals at the time of the study. Data linkage steps followed a previously established methodology [25]. First, the PI RCT data were linked to the DAD using the provincial health number and admission date. Then, DAD variables were used to connect these data with SCM.

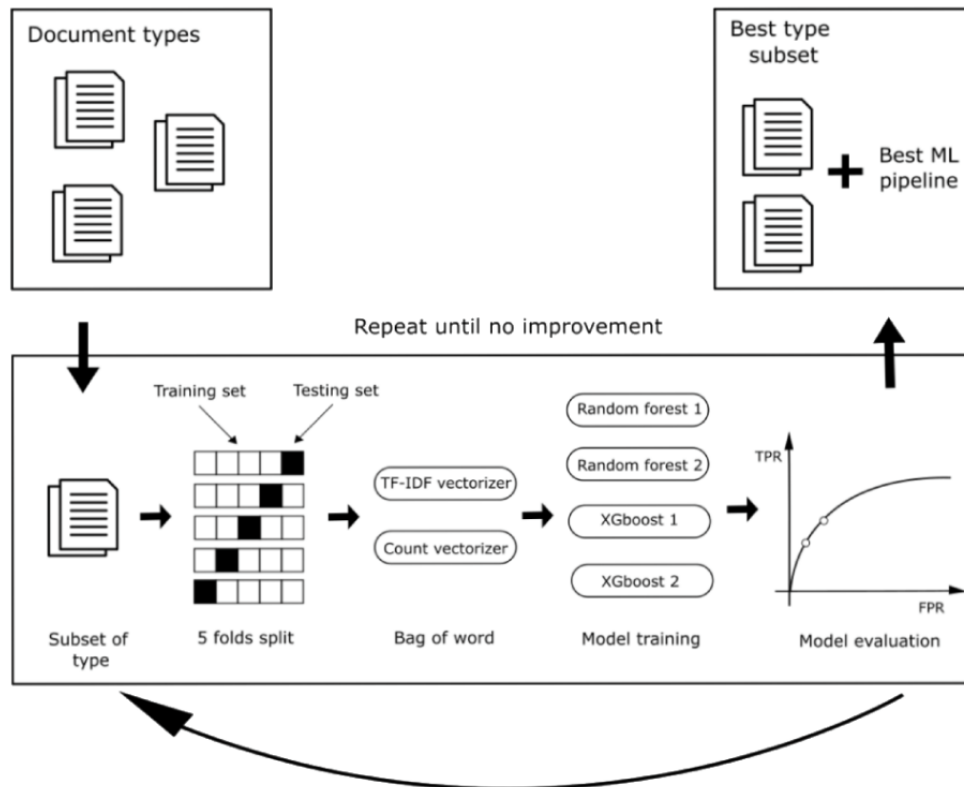
Document Types and Sequential Forward Type Feature Selection

In total, 37 types of documents were noted for the included patients during the clinical trial. Nursing notes were the primary source of suitable HAPI information and constituted the largest proportion of the documents. Among the nursing notes, “Patient Assessment” contained the assessment of skin and wounds under the Integument section. The Integument section described skin integrity, bruises, wound formation, and exposure to air. The “Patient Assessment Neuro” document included the patient’s neurological state, where the main components related to PIs were level of consciousness, communication, and sensory deficit. The “Patient Care” document included patients’ hygiene, activity, exercise, and nutrition, such as mobility, positioning,

and assistance with a meal. The remaining document types contained daily intake and output, physiological indicators, pain scale, and other related data. Discharge summaries, unit transfer notes, and inpatient triage reports were not written for most patients during the clinical trial because the trial was primarily conducted in the middle of the hospital stay.

Forward feature selection was used to determine the best combination of documents with 2 machine learning (ML) models: extreme gradient boosting (XGBoost) and random forest (RF) [26,27]. Forward feature selection is an iterative way to obtain the best subset of features [28]. The analyses began with no feature in the input of models. Then, in each iteration, new features were added and observed for improvements (Figure 3). The experiments were run with each feature from the list of all possible features, where the best predictor was then added to our feature set. This iteration ended when introducing a new feature did not significantly improve the targeted metric. In our experiments, the forward feature selection was performed for every document type. Instead of adding 1 feature in each iteration, all documents belonging to 1 type were added to the input of models. This feature selection stopped when adding a new document type did not increase the target metric. Due to the long convergence time, the forward feature selection was not conducted during the development of the deep learning model. Rather, the same optimal document set determined by ML models was used.

Figure 3. A visual illustration of the sequential forward selection process for identifying feature subsets that maximize the performance of the ML pipeline. The candidate feature subsets were evaluated by using 5-fold cross-validation. For each subset of document types, 40 experiments were conducted with all possible combinations of 5 folds, 2 vectorizers, and 4 ML models. FPR: false positive rate; ML: machine learning; TF-IDF: term frequency-inverse document frequency; TPR: true positive rate.



Natural Language Processing

Bag of Words Preprocessing and ML

All nursing notes of selected document types were merged into 1 text and converted into a bag-of-words (BOW) vector with the count of words or term frequency-inverse document frequency (TF-IDF) vectorizer by using a Python scikit-learn ML library [29-31].

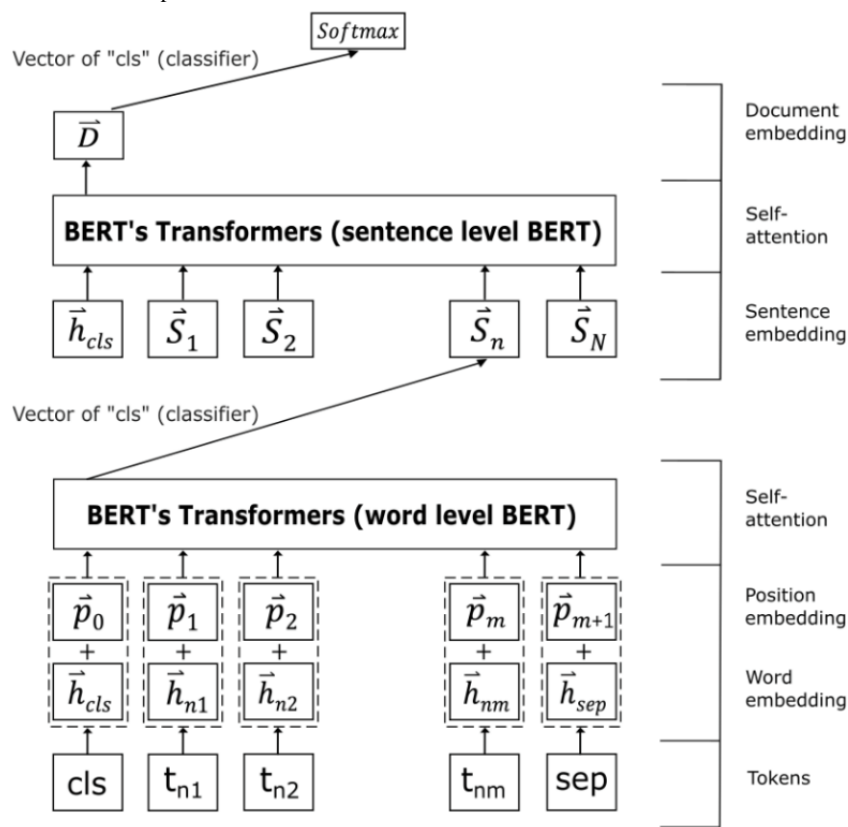
A binary classification model was developed to identify HAPI cases by considering all patients who developed any stage of PI during a hospital stay as positive cases and patients without PIs as the negative cohort. The BOW matrices were used as the independent input for the models. RF and XGBoost classification models were trained to perform classification. These 2 models were chosen because they were representative of ensemble models: RF for bagging and XGboost for boosting. Ensemble models have been shown to display superior performance than a single classifier [32]. Two sets of hyperparameters were tried for each model. The 5-fold cross-validation was conducted to determine the most useful document types, high-performing ML model, and its hyperparameters.

Deep Learning Model

A hierarchical attention network (HAN) structure with bidirectional encoder representations from transformers (BERT) was used to classify the text in the EMR clinical notes [33,34]. BERT is a contextualized word representation model that uses a masked language model that predicts randomly masked words

in a context sequence. Publicly released BERT parameters are trained on corpora such as Wikipedia, which is formatted differently from clinical text. As such, ClinicalBERT, a language model specifically pretrained using clinical notes, was used for the text evaluation [35]. Medical language has been demonstrated to contain vast amounts of discipline-specific jargon, abbreviations, and acronyms while being a domain-specific and technical language [36]. Multiple studies have demonstrated that ClinicalBERT performs better than BERT [37,38]. Therefore, the decision to proceed with ClinicalBERT for our study was made. The ClinicalBERT embedding was not fine-tuned with clinical notes data due to a moderate sample size. Rather, the ClinicalBERT was downloaded and tested from a GitHub repo found in the study where Alzentzer et al observed performance improvements on three common clinical NLP tasks after training BERT models on clinical notes and discharge summaries [38,39]. The document embedding layer weights were not taken from ClinicalBERT. As the maximum sequence length of BERT limits it from handling text with more than 512 tokens, sentence embeddings generated by ClinicalBERT are fed to another transformer to obtain the “document embedding,” a highly abstracted vector capturing global information about the whole document [35,38]. HAPI status was classified based on this document embedding (Figure 4). The project-specific document embedding transformer was trained from the ground up through random initialization. Details of implementation and training of our HAN-BERT models are described in Figure S1 in Multimedia Appendix 1 [40,41].

Figure 4. Composition of input sequence representations for text classification using BERT. The meaning of one sentence is summarized into the vector of [CLS] (classifier), an artifact token concatenated at the beginning of each sentence to become the sentence embedding. The sentence embedding is then fed to another transformer to generate the document embedding. An output layer with SoftMax activation provides the probability of text classification. BERT: bidirectional encoder representations from transformers.



Model Evaluation

We used 5-fold stratified cross-validation to split the 97 positive cases and 183 patient controls into 5 groups. Due to the fact both numbers were not divisible by 5, there was a minor difference in the distribution of cases and controls between groups, although the effort was placed to retain the most similar distribution between the 5 groups. Each time we selected 4 groups as a training set, the remaining group was used as a test set. The splitting was the same for ML and deep learning experiments. A comparison of different document type subsets was executed with the best model to determine which document type subset would yield the best performance of PI detection. To fairly compare our method with ICD-based PI identification algorithms, the classification threshold was tuned to achieve similarly estimated specificities (0.988 and 0.959) of 2 ICD-based algorithms developed and validated in a previous study by Ho et al [1]. The first case definition is more specific and yields greater detection precision. The second definition is more inclusive of nonspecific codes for wounds and is likely to capture a larger number of cases. Sensitivity, specificity, positive predictive value (PPV), negative predictive value (NPV), and F_1 -scores were calculated as target metrics. Additionally, PPV, NPV, sensitivity, and specificity of the 4 algorithms and 2 ICD algorithms were calculated with changing thresholds ranging from 0.05 to 0.95. F_1 -score is a measure of accuracy through a combination of sensitivity and PPV. F_1 has

a maximum score of 1 when both sensitivity and PPV are 1, and a minimum of 0 when either sensitivity or PPV is 0. We calculated the binomial proportion CIs for sensitivity, specificity, PPV, and NPV using the Statsmodels package in Python (Python Software Foundation) [42]. The CIs of the F_1 -score were from 5-fold cross-validation.

Ethics Approval

The study was approved by the Conjoint Health Research Ethics Board at the University of Calgary, Calgary, Alberta (REB13-0794).

Results

Characteristics of Study Cohort

The study included 280 eligible participants (Figure 2). Among the 280 patients, a research nurse identified 183 patients with no HAPIs, and 97 patients were found to have HAPIs. Table 1 provides demographic details of the patient cohort. The P values were calculated with MedCalc's statistical calculators [43,44]. The median age was 68 (IQR 55-79) years. The cohort consisted of 127 (45.36%) females, and the median length of stay was 46 (IQR-79) days. A more detailed review of input data and linguistic inquiry and word count analysis for the number of words, sentences, and patients based on document types can be found in Table S1 in Multimedia Appendix 2.

Table 1. Descriptive statistics of patients (N=280).

Characteristics	All	Patients with pressure injury (n=97)	Patients without pressure injury (n=183)	P value
Female, n (%)	127 (45.36)	42 (43.30)	85 (46.45)	.62
Age (years), median (IQR)	68 (55-79)	68 (59-80)	67 (53-79)	.10
Length of stay (days), median (IQR)	46 (22-104)	48 (28-96)	46 (19-109)	.37
Cerebrovascular disease, n (%)	137 (48.93)	50 (51.55)	87 (47.54)	.52
Chronic obstructive pulmonary disease, n (%)	54 (19.29)	26 (26.80)	28 (15.30)	.02
Congestive heart failure, n (%)	66 (23.57)	30 (30.93)	36 (19.67)	.04
Myocardial infarction, n (%)	33 (11.79)	9 (9.28)	24 (13.11)	.34
Dementia, n (%)	27 (9.64)	10 (10.31)	17 (9.29)	.79
Peripheral vascular disease, n (%)	51 (18.21)	21 (21.65)	30 (16.39)	.28
Hemiplegia or paraplegia, n (%)	119 (42.50)	39 (40.21)	80 (43.72)	.57
Leukemia, n (%)	2 (0.71)	1 (1.03)	1 (0.55)	.62
Lymphoma, n (%)	4 (1.43)	2 (2.06)	2 (1.09)	.50
Peptic ulcer disease, n (%)	41 (14.64)	21 (21.65)	20 (10.93)	.02
Moderate or severe renal disease, n (%)	55 (19.64)	28 (28.87)	27 (14.75)	.005
Liver disease, n (%)	23 (8.21)	8 (8.25)	15 (8.20)	.99
Diabetes mellitus, n (%)	91 (32.50)	39 (40.21)	52 (28.42)	.045
Solid tumor, n (%)	50 (17.86)	19 (19.59)	31 (16.94)	.58
Connective tissue, n (%) disease	51 (18.21)	24 (24.74)	27 (14.75)	.04
History of smoking, n (%)	118 (42.14)	45 (46.39)	73 (39.89)	.30
Currently smoking, n (%)	37 (13.21)	11 (11.34)	26 (14.21)	.50
History of illicit drug use, n (%)	15 (5.36)	7 (7.22)	8 (4.37)	.32
Currently use illicit drugs, n (%)	8 (2.86)	3 (3.09)	5 (2.73)	.85

Data Linkage and Extraction

Table 2 shows the patient and document count and document word count for the patients eligible for this study. Most PI-positive patients had “Patient assessment” document type

(60 (61.86%) patients with HAPI versus 82 (44.81%) patients without HAPI), and patients in the negative groups predominantly had “Patient assessment Neuro.” However, patients from both groups had a similar amount of “Patient care” during the trial (Table 2).

Table 2. Characteristics of extracted documents, different components of nursing notes, and the average number of documents written by nurses.

Document type	All (N=280)	Patients with HAPI (n=97)	Patients without HAPI (n=183)
Patient assessment			
Number of patients with this type of document, n (%)	142 (50.71)	60 (61.86)	82 (44.81)
Number of notes per patient, median (IQR)	1.00 (0.00-17.00)	12.00 (0.00-18.00)	0.00 (0.00-16.00)
Word count per note, median (IQR)	376.00 (306.00-430.00)	385.00 (311.00-441.00)	370.00 (303.00-421.00)
Patient assessment neuro			
Number of patients with this type of document, n (%)	141 (50.36)	39 (40.21)	102 (55.74)
Number of notes per patient, median (IQR)	2.50 (0.00-13.00)	0.00 (0.00-13.00)	8.00 (0.00-13.00)
Word count per note, median (IQR)	428.00 (343.0-498.0)	424.00 (324.00-493.50)	430.00 (346.00-499.00)
Patient care			
Number of patients with this type of document, n (%)	280 (100)	97 (100)	183(100)
Number of notes per patient, median (IQR)	16.00 (13.00-21.20)	16.00 (13.00-19.00)	16.00 (13.00-23.50)
Word count per note, median (IQR)	138.00 (72.00-179.00)	147.00 (91.00-185.00)	133.00 (66.00-176.00)

Document Subset and Classification Models

Across a subset of document types and all tested classification techniques, the combination of Patient Assessment, Patient Assessment Neuro, and Patient Care yielded the highest outputs in terms of target metrics.

The TF-IDF vectorizer with RF classifier demonstrated the best performance in terms of sensitivity, PPV, and NPV when fixed at the specificity of 0.988 and 0.959 thresholds. The performance results are reported in [Table 3](#).

For a specificity of 0.988, the sensitivity of the (TF-IDF+RF) EMR-based model was 0.464 (95% CI 0.365-0.563), which surpassed the sensitivity 0.277 (95% CI 0.174-0.380) achieved by the ICD-based algorithm [1]. The PPV of our model had a mean of 0.938 (95% CI 0.869-1.000), which is higher than the reported 0.917 (95% CI 0.854-0.980) of the ICD algorithm. The NPV was 0.776 (95% CI 0.722-0.830), which is also higher than the 0.739 (95% CI 0.638-0.840) reported in the ICD validation [1]. For a specificity of 0.959 achieved by the loose ICD-based algorithm, the EMR model sensitivity reached 0.546 (95% CI 0.447-0.645) compared to 0.328 (95% CI 0.220-0.436) found in ICD reporting [1]. Both PPV and NPV of EMR model

were also higher (0.855 (95% CI 0.767-0.943) vs 0.793 (95% CI 0.700-0.886) and 0.798 (95% CI 0.745-0.851 vs 0.746 (95% CI 0.646-0.846) than those detected by ICD algorithm respectively. The deep learning model underperformed with the area under the receiver operating characteristic curve (AUC-ROC) score of 0.68 (SD 0.04), compared to the RF with the highest area under the curve (AUC) score of 0.80 (SD 0.08), followed by XGBoost with the AUC score of 0.75 (SD 0.07; [Figure 5](#)). Considering the prevalence of 34.6% in our study, the baseline area under the precision-recall curve (AU-PRC) is 0.346. [Figure 6](#) shows that an AU-PRC of 0.77 (SD 0.06) was achieved for the RF models using TF-IDF tokenization, 0.74 (SD 0.08) was achieved for the RF models using count tokenization, 0.67 (SD 0.04) was achieved for the XGBoost models, and 0.60 (SD 0.06) for the deep learning models. These results are greater than 0.346, and we conclude that these classifiers do not discriminate by random chance and perform well in finding positive HAPI cases without accidentally marking negative patients as positive. [Figure S1](#) in [Multimedia Appendix 3](#) shows the PPV, NPV, sensitivity, and specificity of the 4 algorithms and 2 ICD algorithms, with changing thresholds ranging between 0.05 and 0.95.

Table 3. The performance of NLP^a methods on free-text electronic medical record documents at varying thresholds for the probability of pressure injury detection. The model was compared to ICD^b algorithms such that the model was trained to mimic its specificity.

Model	Sensitivity % (95% CI)	Specificity % (95% CI)	PPV ^c % (95% CI)	NPV ^d % (95% CI)	F ₁ -score % (95% CI)
Specificity near 0.988					
ICD (Ho et al [1])	0.277 (0.174-0.380)	0.988 (0.963-1.013)	0.917 (0.854-0.980)	0.739 (0.638-0.840)	0.425 (0.312-0.538)
TF-IDF ^e +random forest ^f	0.464 (0.365-0.563)	0.984 (0.965-1.000)	0.938 (0.869-1.000)	0.776 (0.722-0.830)	0.612 (0.473-0.751)
Count+random forest	0.412 (0.314-0.510)	0.978 (0.957-0.999)	0.909 (0.824-0.994)	0.758 (0.704-0.813)	0.550 (0.361-0.739)
TF-IDF+XGBoost ^g	0.309 (0.217-0.401)	0.973 (0.949-0.996)	0.857 (0.741-0.973)	0.727 (0.671-0.782)	0.450 (0.340-0.559)
Word Embedding+BERT ^h	0.268 (0.180-0.356)	0.978 (0.957-0.999)	0.867 (0.745-0.988)	0.716 (0.660-0.772)	0.394 (0.207-0.580)
Specificity near 0.959					
ICD (Ho et al [1])	0.328 (0.220-0.436)	0.959 (0.913-0.100)	0.793 (0.700-0.886)	0.746 (0.646-0.846)	0.464 (0.350-0.578)
TF-IDF+random forest	0.546 (0.447-0.645)	0.951 (0.919-0.982)	0.855 (0.767-0.943)	0.798 (0.745-0.851)	0.665 (0.577-0.753)
Count+random forest	0.423 (0.324-0.521)	0.956 (0.927-0.986)	0.837 (0.733-0.940)	0.758 (0.702-0.813)	0.546 (0.359-0.733)
TF-IDF+XGBoost	0.423 (0.324-0.521)	0.956 (0.927-0.986)	0.837 (0.733-0.940)	0.758 (0.702-0.813)	0.552 (0.404-0.699)
Word embedding+BERT	0.289 (0.198-0.379)	0.967 (0.941-0.993)	0.824 (0.695-0.952)	0.720 (0.663-0.776)	0.420 (0.280-0.560)

^aNLP: natural language processing.

^bICD: International Classification of Diseases.

^cPPV: positive predictive value.

^dNPV: negative predictive value.

^eTF-IDF: term frequency-inverse document frequency.

^fThe best model.

^gXGBoost: extreme gradient boosting.

^hBERT: bidirectional encoder representations from transformers.

Figure 5. The ROC curves derived from the random forest with TF-IDF and word count, XGBoost, and deep learning models. AUC: area under the curve; ROC: receiver operating characteristic; TF-IDF: term frequency-inverse document frequency; XGBoost: eXtreme gradient boosting.

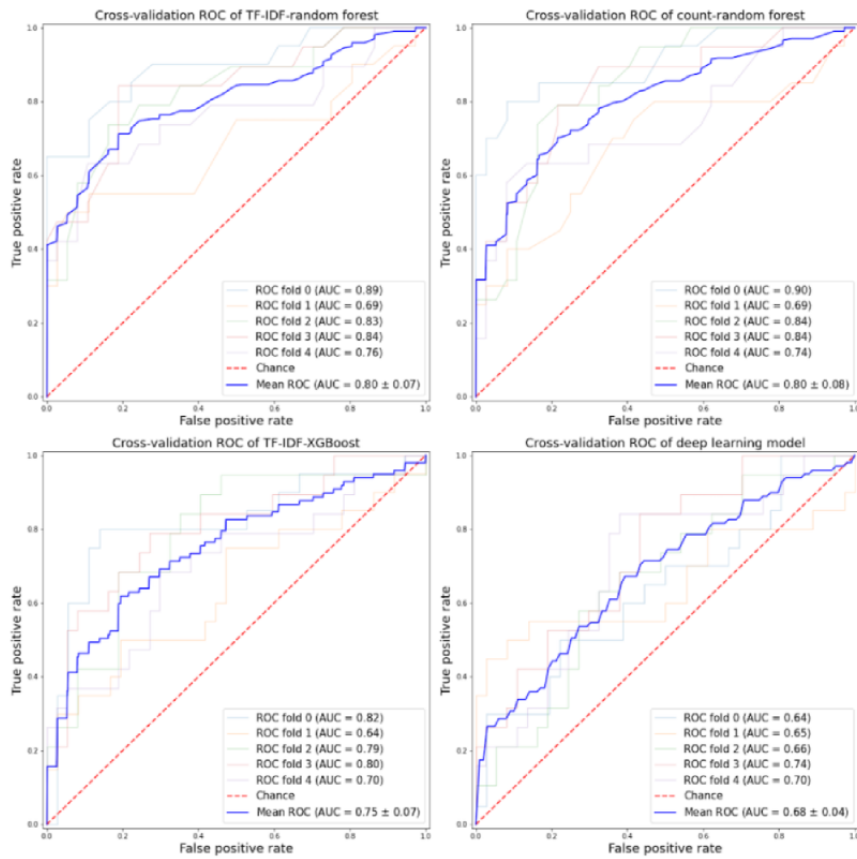
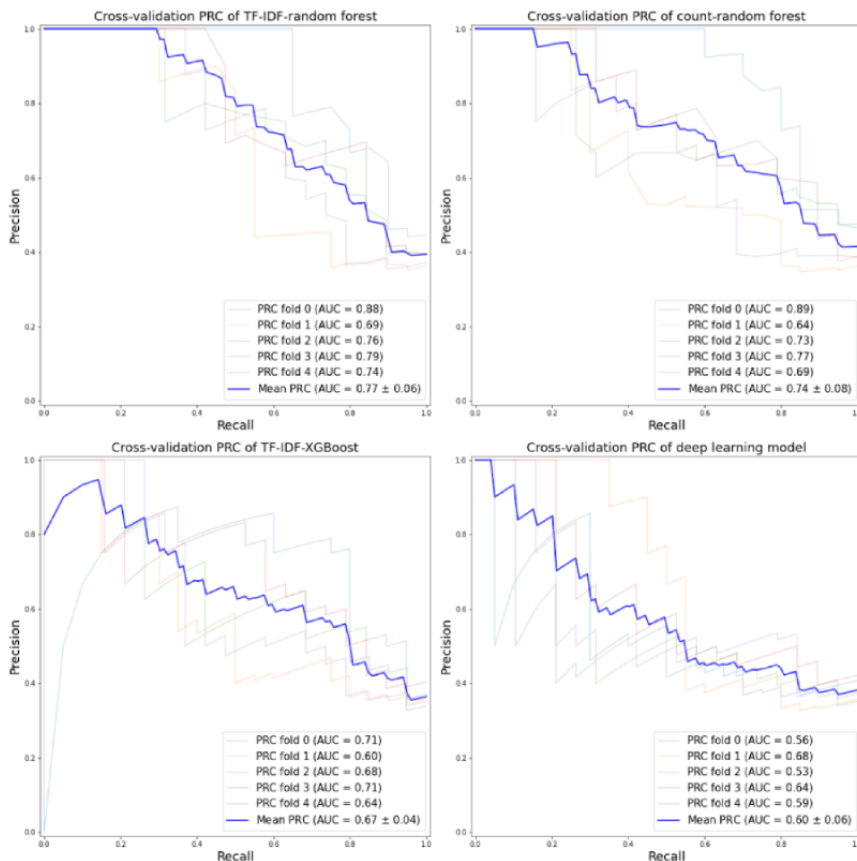


Figure 6. The area under the precision-recall curve (AU-PRC) performance of 4 models: random forest with TF-IDF and word count, XGBoost, deep learning model. AUC: area under the curve; PRC: precision-recall curve; TF-IDF: term frequency-inverse document frequency; XGBoost: eXtreme gradient boosting;



Discussion

Principal Findings

Multiple methods were applied, and different combinations of clinical text were analyzed to determine the efficiency of NLP models in detecting HAPIs from nursing notes. The results of NLP models were compared with the ICD-based algorithm reported in the previous study [1]. An AUC of 0.80 (SD 0.07) of the ML model indicates fair accuracy in terms of produced sensitivity and specificity. The results demonstrate that different combinations of EMR data leverage NLP models to improve upon ICD-10-based HAPI case definitions. The TF-IDF with RF produced higher sensitivity at a strict specificity level. The satisfactory performance of ML models indicates that the free-text documented during hospitalization contains valuable information for HAPI detection. Developing algorithms using EMR data will facilitate the timely and accurate capture of HAPI incidences and measure the quality of nursing practice during patient hospitalization.

From the forward document type selection, we found that apart from the notes that directly document skin conditions in the patient assessment, the entries noting the patient's consciousness, nutrition, and mobility were helpful in indicating HAPI. This makes clinical sense because the reduced level of consciousness, nutrition, and mobility are factors that may contribute to HAPI [45]. In addition, our findings align with several risk factors of the Braden Scale [46]. Given that many factors of the Braden Scale are documented in nursing notes daily, it may be promising to use NLP to automatically extract the Braden Scale's factors and achieve better PI detection or prediction upon the automatically rated Braden Scale [46].

The results showed that the XGBoost and RF methods perform better than the advanced deep learning models by a large margin. The joint effort of the TF-IDF vectorizer and tree-based classifier enabled the pipeline to stay robust to irrelevant vocabularies, even when the sample size was smaller than the feature size. The feature selection played a role in this task because a great part of the text in clinical documents was not relevant to HAPIs and only contributed noise for the classification task. On the other hand, deep learning models allowed every input word to contribute to the document embedding upon which the model judged the presence of HAPIs. The suboptimal performance of the deep learning model may have been avoided if the transformers' attention mechanisms had more training samples to converge. The noisy data and not a very large sample size were possibly the main factors that made the deep learning models perform poorly. However, this hypothesis needs further examination.

Compared to these previous studies that used EMR to automate phenotyping, our model achieved higher sensitivity while reporting comparable values for performance metrics such as PPV and NPV. Furthermore, our model can identify HAPIs with high specificity and improved sensitivity during the first three days in routine clinical practice settings. Melton et al [47] found NLP to be reliable and effective in detecting 16 out of 65 adverse events in 1000 manually reviewed charts. The model by Melton et al [47] then processed all inpatient cases with EMR

discharge summaries, achieving high specificity (0.9996) and low sensitivity (0.28). Our model results are in line with other studies that used free-text clinical notes to predict incidences of distinct adverse events [48-51].

Limitations

The study is not without limitations. First, the exclusion of ICU patients due to data elements being distinct from SCM led to a smaller sample size and a narrower clinical cohort. Nevertheless, the remaining data from the clinical trial represented a population at risk for HAPIs. Second, both the patient and nurse knew at the admission of a clinical trial measuring PI would be the trial goal, which may have impacted the data entry quality of PI and frequency of patients to report PI-related discomfort. Third, the model produced relatively modest sensitivity. However, this sensitivity is deemed valuable, given that the specificity was set to a high threshold, and the input was restricted to the first 72 hours after enrollment [16]. The sensitivity reported in similar studies used the whole or more extended hospitalization stay and more data elements [50-52]. In addition, the sensitivity of our study was obtained through a comparison to a clinical trial instead of chart review data. Chart review does not always capture all positive cases due to possible errors in the review process [53,54]. Fourth, the comparison with deep learning is not likely to be very fair because BERT-based models are usually applied to larger cohorts. Nevertheless, our result can be served as a reference for model selection for researchers working with similar sample sizes. Prabhakar et al applied ClinicalBERT to phenotype 10 diseases on a cohort consisting of 1610 discharge summaries [41]. When only using ClinicalBERT, they obtained a very similar F_1 -score (0.46) compared to our result [41]. The suboptimal performances of the advanced deep learning model may suggest that study needs to be more evolved before applying deep learning to free-text-based clinical phenotyping. Tree-based ML models are recommended for detecting adverse event conditions from noisy, moderately sized text samples.

Future Directions

The present work focused on demonstrating ML models on cross-sectional EMR data can outperform the ICD-based PI identification algorithm. Future directions could include (1) leveraging cost-sensitive learning to assign various weights to assess the impact of misclassifying the patients with a PI, (2) identification of the potential risk features or predictors that may be associated with PI, (3) comparison of HAN-BERT against other novel NN structures, and (4) detailed ablation studies for assessing the performance of components on the designed models that will hopefully be integrated into a clinical decision-support system. These studies will require larger sample sizes than our current pilot study, but our current work can be used to create such a cohort.

Conclusions

Our study revealed the feasibility of using inpatient clinical notes documented for 3 days to detect HAPIs with increased accuracy over ICD methods. NLP and ML application on inpatient clinical notes allowed better and more timely use of the clinical narratives compared to summarizing them into ICD

codes and DAD, thereby being a promising solution for precise, time-sensitive, population-based disease phenotyping. With the advent of digital technologies in health care, the results contribute toward an automated approach to better cohort identification, patient surveillance, and quality improvement for the treatment of hospital-acquired adverse events. The

application of the model is particularly relevant for effectively mining clinical data that does not capture a large sample size for adverse effects phenotyping. The proposed method of identifying patients in acute care hospitals who are likely to have or develop PI will most likely be used by front-line hospital staff to prevent or manage PI earlier and more effectively.

Conflicts of Interest

None declared.

Multimedia Appendix 1

Implementation details.

[\[DOCX File , 84 KB - ai_v2i1e41264_app1.docx \]](#)

Multimedia Appendix 2

Linguistic inquiry and word count analysis.

[\[DOCX File , 16 KB - ai_v2i1e41264_app2.docx \]](#)

Multimedia Appendix 3

Positive predictive value (PPV), negative predictive value (NPV), sensitivity, and specificity of the four algorithms and two International Classification of Diseases (ICD) algorithms, with changing thresholds ranging between 0.05 and 0.95.

[\[DOCX File , 103 KB - ai_v2i1e41264_app3.docx \]](#)

References

1. Ho C, Jiang J, Eastwood CA, Wong H, Weaver B, Quan H. Validation of two case definitions to identify pressure ulcers using hospital administrative data. *BMJ Open* 2017 Aug 28;7(8):e016438 [[FREE Full text](#)] [doi: [10.1136/bmjopen-2017-016438](https://doi.org/10.1136/bmjopen-2017-016438)] [Medline: [28851785](https://pubmed.ncbi.nlm.nih.gov/28851785/)]
2. Woodbury MG, Houghton PE. Prevalence of pressure ulcers in Canadian healthcare settings. *Ostomy Wound Manage* 2004 Oct;50(10):22-24, 26, 28, 30, 32, 34, 36 [[FREE Full text](#)] [Medline: [15509880](https://pubmed.ncbi.nlm.nih.gov/15509880/)]
3. Lyder CH, Wang Y, Metersky M, Curry M, Kliman R, Verzier NR, et al. Hospital-acquired pressure ulcers: results from the national Medicare Patient Safety Monitoring System study. *J Am Geriatr Soc* 2012 Sep;60(9):1603-1608. [doi: [10.1111/j.1532-5415.2012.04106.x](https://doi.org/10.1111/j.1532-5415.2012.04106.x)] [Medline: [22985136](https://pubmed.ncbi.nlm.nih.gov/22985136/)]
4. Ebi WE, Hirko GF, Mijena DA. Nurses' knowledge to pressure ulcer prevention in public hospitals in Wollega: a cross-sectional study design. *BMC Nurs* 2019;18:20 [[FREE Full text](#)] [doi: [10.1186/s12912-019-0346-y](https://doi.org/10.1186/s12912-019-0346-y)] [Medline: [31139012](https://pubmed.ncbi.nlm.nih.gov/31139012/)]
5. Baharestani MM, Black JM, Carville K, Clark M, Cuddigan JE, Dealey C, et al. Dilemmas in measuring and using pressure ulcer prevalence and incidence: an international consensus. *Int Wound J* 2009;6(2):97-104 [[FREE Full text](#)] [doi: [10.1111/j.1742-481X.2009.00593.x](https://doi.org/10.1111/j.1742-481X.2009.00593.x)] [Medline: [19432659](https://pubmed.ncbi.nlm.nih.gov/19432659/)]
6. O'Malley KJ, Cook KF, Price MD, Wildes KR, Hurdle JF, Ashton CM. Measuring diagnoses: ICD code accuracy. *Health Serv Res* 2005;40(5 Pt 2):1620-1639 [[FREE Full text](#)] [doi: [10.1111/j.1475-6773.2005.00444.x](https://doi.org/10.1111/j.1475-6773.2005.00444.x)] [Medline: [16178999](https://pubmed.ncbi.nlm.nih.gov/16178999/)]
7. Dutta B, Hwang HG. The adoption of electronic medical record by physicians: a PRISMA-compliant systematic review. *Medicine (Baltimore)* 2020;99(8):e19290 [[FREE Full text](#)] [doi: [10.1097/MD.0000000000019290](https://doi.org/10.1097/MD.0000000000019290)] [Medline: [32080145](https://pubmed.ncbi.nlm.nih.gov/32080145/)]
8. Xu Y, Lee S, Martin E, D'souza AG, Doktorchik CTA, Jiang J, et al. Enhancing ICD-code-based case definition for heart failure using electronic medical record data. *J Card Fail* 2020;26(7):610-617. [doi: [10.1016/j.cardfail.2020.04.003](https://doi.org/10.1016/j.cardfail.2020.04.003)] [Medline: [32304875](https://pubmed.ncbi.nlm.nih.gov/32304875/)]
9. Ford E, Carroll JA, Smith HE, Scott D, Cassell JA. Extracting information from the text of electronic medical records to improve case detection: a systematic review. *J Am Med Inform Assoc* 2016;23(5):1007-1015 [[FREE Full text](#)] [doi: [10.1093/jamia/ocv180](https://doi.org/10.1093/jamia/ocv180)] [Medline: [26911811](https://pubmed.ncbi.nlm.nih.gov/26911811/)]
10. Lin WC, Chen JS, Chiang MF, Hribar MR. Applications of artificial intelligence to electronic health record data in ophthalmology. *Transl Vis Sci Technol* 2020;9(2):13 [[FREE Full text](#)] [doi: [10.1167/tvst.9.2.13](https://doi.org/10.1167/tvst.9.2.13)] [Medline: [32704419](https://pubmed.ncbi.nlm.nih.gov/32704419/)]
11. Bao XY, Huang WJ, Zhang K, Jin M, Li Y, Niu CZ. [A customized method for information extraction from unstructured text data in the electronic medical records]. *Beijing Da Xue Xue Bao Yi Xue Ban* 2018;50(2):256-263 [[FREE Full text](#)] [doi: [10.3969/j.issn.1671-167X.2018.02.010](https://doi.org/10.3969/j.issn.1671-167X.2018.02.010)] [Medline: [29643524](https://pubmed.ncbi.nlm.nih.gov/29643524/)]
12. Juhn Y, Liu H. Artificial intelligence approaches using natural language processing to advance EHR-based clinical research. *J Allergy Clin Immunol* 2020;145(2):463-469 [[FREE Full text](#)] [doi: [10.1016/j.jaci.2019.12.897](https://doi.org/10.1016/j.jaci.2019.12.897)] [Medline: [31883846](https://pubmed.ncbi.nlm.nih.gov/31883846/)]

13. Szlosek DA, Ferrett J. Using machine learning and natural language processing algorithms to automate the evaluation of clinical decision support in electronic medical record systems. *EGEMS (Wash DC)* 2016;4(3):1222 [FREE Full text] [doi: [10.13063/2327-9214.1222](https://doi.org/10.13063/2327-9214.1222)] [Medline: [27683664](https://pubmed.ncbi.nlm.nih.gov/27683664/)]
14. Maarseveen TD, Meinderink T, Reinders MJT, Knitza J, Huizinga TWJ, Kleyer A, et al. Machine learning electronic health record identification of patients with rheumatoid arthritis: algorithm pipeline development and validation study. *JMIR Med Inform* 2020;8(11):e23930 [FREE Full text] [doi: [10.2196/23930](https://doi.org/10.2196/23930)] [Medline: [33252349](https://pubmed.ncbi.nlm.nih.gov/33252349/)]
15. Wong J, Horwitz MM, Zhou L, Toh S. Using machine learning to identify health outcomes from electronic health record data. *Curr Epidemiol Rep* 2018;5(4):331-342 [FREE Full text] [doi: [10.1007/s40471-018-0165-9](https://doi.org/10.1007/s40471-018-0165-9)] [Medline: [30555773](https://pubmed.ncbi.nlm.nih.gov/30555773/)]
16. Xu H, Fu Z, Shah A, Chen Y, Peterson NB, Chen Q, et al. Extracting and integrating data from entire electronic health records for detecting colorectal cancer cases. *AMIA Annu Symp Proc* 2011;2011:1564-1572 [FREE Full text] [Medline: [22195222](https://pubmed.ncbi.nlm.nih.gov/22195222/)]
17. Canada's health care system. Government of Canada. Canada; 2019. URL: <https://www.canada.ca/en/health-canada/services/health-care-system/reports-publications/health-care-system/canada.html> [accessed 2022-12-16]
18. Denny K, Lawand C, Perry SD. Compromised wounds in Canada. *Healthc Q* 2014;17(1):7-10. [doi: [10.12927/hcq.2014.23787](https://doi.org/10.12927/hcq.2014.23787)] [Medline: [24844713](https://pubmed.ncbi.nlm.nih.gov/24844713/)]
19. About CIHI. Canadian Institute for Health Information. URL: <https://www.cihi.ca/en/about-cihi> [accessed 2023-01-24]
20. Discharge abstract database metadata (DAD). Canadian Institute for Health Information. URL: <https://www.cihi.ca/en/discharge-abstract-database-metadata-dad> [accessed 2023-01-24]
21. Wong H, Kaufman J, Baylis B, Conly JM, Hogan DB, Stelfox HT, et al. Efficacy of a pressure-sensing mattress cover system for reducing interface pressure: study protocol for a randomized controlled trial. *Trials* 2015 Sep 29;16:434 [FREE Full text] [doi: [10.1186/s13063-015-0949-x](https://doi.org/10.1186/s13063-015-0949-x)] [Medline: [26420303](https://pubmed.ncbi.nlm.nih.gov/26420303/)]
22. Edsberg LE, Black JM, Goldberg M, McNichol L, Moore L, Sieggreen M. Revised National Pressure Ulcer Advisory Panel pressure injury staging system: revised pressure injury staging system. *J Wound Ostomy Continence Nurs* 2016;43(6):585-597 [FREE Full text] [doi: [10.1097/WON.0000000000000281](https://doi.org/10.1097/WON.0000000000000281)] [Medline: [27749790](https://pubmed.ncbi.nlm.nih.gov/27749790/)]
23. Black J, Baharestani M, Cuddigan J, Dorner B, Edsberg L, Langemo D, National Pressure Ulcer Advisory Panel. National Pressure Ulcer Advisory Panel's updated pressure ulcer staging system. *Dermatol Nurs* 2007 Aug;19(4):343-349; quiz 350. [Medline: [17874603](https://pubmed.ncbi.nlm.nih.gov/17874603/)]
24. Lee S, Xu Y, D Apos Souza AG, Martin EA, Doktorchik C, Zhang Z, et al. Unlocking the potential of electronic health records for health research. *Int J Popul Data Sci* 2020 Jan 30;5(1):1123 [FREE Full text] [doi: [10.23889/ijpds.v5i1.1123](https://doi.org/10.23889/ijpds.v5i1.1123)] [Medline: [32935049](https://pubmed.ncbi.nlm.nih.gov/32935049/)]
25. Lee S, Li B, Martin EA, D'Souza AG, Jiang J, Doktorchik C, et al. CREATE: a new data resource to support cardiac precision health. *CJC Open* 2021 May;3(5):639-645 [FREE Full text] [doi: [10.1016/j.cjco.2020.12.019](https://doi.org/10.1016/j.cjco.2020.12.019)] [Medline: [34036259](https://pubmed.ncbi.nlm.nih.gov/34036259/)]
26. Chen T, Guestrin C. XGBoost: a scalable tree boosting system. New York, NY, USA: Association for Computing Machinery; 2016 Presented at: KDD '16: The 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining; August 13-17, 2016; San Francisco, California p. 785-794 URL: <https://dl.acm.org/doi/10.1145/2939672.2939785> [doi: [10.1145/2939672.2939785](https://doi.org/10.1145/2939672.2939785)]
27. Ho TK. Random decision forests. 1995 Presented at: Proceedings of 3rd International Conference on Document Analysis and Recognition; August 14-16, 1995; Montreal, QC, Canada p. 278-282 URL: <https://ieeexplore.ieee.org/document/598994> [doi: [10.1109/icdar.1995.598994](https://doi.org/10.1109/icdar.1995.598994)]
28. Last M, Kandel A, Maimon O. Information-theoretic algorithm for feature selection. *Pattern Recogn Lett* 2001 May;22(6-7):799-811 [FREE Full text] [doi: [10.1016/s0167-8655\(01\)00019-8](https://doi.org/10.1016/s0167-8655(01)00019-8)]
29. Harris ZS. Distributional structure. *WORD* 2015 Dec 04;10(2-3):146-162 [FREE Full text] [doi: [10.1080/00437956.1954.11659520](https://doi.org/10.1080/00437956.1954.11659520)]
30. Sammut C, Webb GI, editors. *Encyclopedia of Machine Learning*. New York: Springer; 2010. [doi: [10.1007/978-0-387-30164-8](https://doi.org/10.1007/978-0-387-30164-8)]
31. Scikit-learn: machine learning in python. Scikit-learn. URL: <https://scikit-learn.org/stable/> [accessed 2023-01-24]
32. Sagi O, Rokach L. Ensemble learning: a survey. *Wiley Interdisciplinary Reviews-Data Mining and Knowledge Discovery* 2018;8(4):e1249 [FREE Full text] [doi: [10.1002/widm.1249](https://doi.org/10.1002/widm.1249)]
33. Yang Z, Yang D, Dyer C, He X, Smola A, Hovy E. Hierarchical attention networks for document classification. : Association for Computational Linguistics; 2016 Presented at: Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies; June 2016; San Diego, California p. 1480-1489 URL: <https://aclanthology.org/N16-1174/> [doi: [10.18653/v1/n16-1174](https://doi.org/10.18653/v1/n16-1174)]
34. Devlin J, Chang MW, Lee K, Toutanova K. BERT: pre-training of deep bidirectional transformers for language understanding. arXiv. Preprint posted online October 11, 2018 2018 [FREE Full text]
35. Huang K, Altosaar J, Ranganath R. ClinicalBERT: modeling clinical notes and predicting hospital readmission. arXiv. Preprint posted online April 10, 2019 2019 [FREE Full text]
36. Hull M. Medical language proficiency: A discussion of interprofessional language competencies and potential for patient risk. *Int J Nurs Stud* 2016 Feb;54:158-172. [doi: [10.1016/j.ijnurstu.2015.02.015](https://doi.org/10.1016/j.ijnurstu.2015.02.015)] [Medline: [25863658](https://pubmed.ncbi.nlm.nih.gov/25863658/)]

37. Lin C, Bethard S, Dligach D, Sadeque F, Savova G, Miller TA. Does BERT need domain adaptation for clinical negation detection? *J Am Med Inform Assoc* 2020 Apr 01;27(4):584-591 [FREE Full text] [doi: [10.1093/jamia/ocaa001](https://doi.org/10.1093/jamia/ocaa001)] [Medline: [32044989](https://pubmed.ncbi.nlm.nih.gov/32044989/)]
38. Alsentzer E, Murphy JR, Boag W, Weng WH, Jindi D, Naumann T, et al. Publicly available clinical BERT embeddings. arXiv. Preprint posted online April 6, 2019 2019 [FREE Full text] [doi: [10.18653/v1/w19-1909](https://doi.org/10.18653/v1/w19-1909)]
39. Alsentzer E, Murphy JR, Boag W, Weng WH, Jindi D, Naumann T, et al. Publicly available clinical BERT embeddings. 2019 Presented at: Proceedings of the 2nd Clinical Natural Language Processing Workshop; June 2019; Minneapolis, Minnesota, USA p. 72-78. [doi: [10.18653/v1/W19-1909](https://doi.org/10.18653/v1/W19-1909)]
40. Liu NF, Gardner M, Belinkov Y, Peters ME, Smith NA. Linguistic Knowledge and Transferability of Contextual Representations. 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 2019; Minneapolis, Minnesota, USA p. 1073-1094. [doi: [10.18653/v1/N19-1112](https://doi.org/10.18653/v1/N19-1112)]
41. Prabhakar A, Shidharth S, Kamath S. Neural language modeling of unstructured clinical notes for automated patient phenotyping. In: 2022 56th Annual Conference on Information Sciences and Systems (CISS). 2022 Presented at: 2022 56th Annual Conference on Information Sciences and Systems (CISS); 9-11 March 2022; Princeton, NJ, USA p. 142-147 URL: <https://ieeexplore.ieee.org/document/9751198>
42. Seabold S, Perktold J. Statsmodels: econometric and statistical modeling with python. 2010 Presented at: Proceedings of the 9th Python in Science Conference; June 28-30, 2010; Austin p. 92-96 URL: <https://conference.scipy.org/proceedings/scipy2010/seabold.html> [doi: [10.25080/majora-92bf1922-011](https://doi.org/10.25080/majora-92bf1922-011)]
43. Comparison of proportions calculator. MedCalc. URL: https://www.medcalc.org/calc/comparison_of_proportions.php [accessed 2023-01-24]
44. Comparison of means calculator. MedCalc. URL: https://www.medcalc.org/calc/comparison_of_means.php [accessed 2023-01-24]
45. Raetz JGM, Wick KH. Common questions about pressure ulcers. *Am Fam Phys* 2015 Nov 15;92(10):888-894 [FREE Full text] [Medline: [26554282](https://pubmed.ncbi.nlm.nih.gov/26554282/)]
46. Braden B, Bergstrom N. A conceptual schema for the study of the etiology of pressure sores. *Rehabil Nurs* 1987;12(1):8-12. [doi: [10.1002/j.2048-7940.1987.tb00541.x](https://doi.org/10.1002/j.2048-7940.1987.tb00541.x)] [Medline: [3643620](https://pubmed.ncbi.nlm.nih.gov/3643620/)]
47. Melton GB, Hripcsak G. Automated detection of adverse events using natural language processing of discharge summaries. *J Am Med Inform Assoc* 2005;12(4):448-457 [FREE Full text] [doi: [10.1197/jamia.M1794](https://doi.org/10.1197/jamia.M1794)] [Medline: [15802475](https://pubmed.ncbi.nlm.nih.gov/15802475/)]
48. Sotoodeh M, Gero ZH, Zhang W, Hertzberg VS, Ho JC. Pressure ulcer injury in unstructured clinical notes: detection and interpretation. *AMIA Annu Symp Proc* 2020;2020:1160-1169 [FREE Full text] [Medline: [33936492](https://pubmed.ncbi.nlm.nih.gov/33936492/)]
49. Levy JJ, Lima JF, Miller MW, Freed GL, O'Malley AJ, Emeny RT. Machine learning approaches for hospital acquired pressure injuries: a retrospective study of electronic medical records. *Front Med Technol* 2022;4:926667 [FREE Full text] [doi: [10.3389/fmedt.2022.926667](https://doi.org/10.3389/fmedt.2022.926667)] [Medline: [35782577](https://pubmed.ncbi.nlm.nih.gov/35782577/)]
50. Kwon O, Na W, Kang H, Jun TJ, Kweon J, Park GM, et al. Electronic medical record-based machine learning approach to predict the risk of 30-day adverse cardiac events after invasive coronary treatment: machine learning model development and validation. *JMIR Med Inform* 2022 May 11;10(5):e26801 [FREE Full text] [doi: [10.2196/26801](https://doi.org/10.2196/26801)] [Medline: [35544292](https://pubmed.ncbi.nlm.nih.gov/35544292/)]
51. Borjali A, Magnéli M, Shin D, Malchau H, Muratoglu OK, Varadarajan KM. Natural language processing with deep learning for medical adverse event detection from free-text medical narratives: a case study of detecting total hip replacement dislocation. *Comput Biol Med* 2021 Feb;129:104140. [doi: [10.1016/j.compbio.2020.104140](https://doi.org/10.1016/j.compbio.2020.104140)] [Medline: [33278631](https://pubmed.ncbi.nlm.nih.gov/33278631/)]
52. Rojas JC, Carey KA, Edelson DP, Venable LR, Howell MD, Churpek MM. Predicting intensive care unit readmission with machine learning using electronic health record data. *Ann Am Thorac Soc* 2018 Jul;15(7):846-853 [FREE Full text] [doi: [10.1513/AnnalsATS.201710-787OC](https://doi.org/10.1513/AnnalsATS.201710-787OC)] [Medline: [29787309](https://pubmed.ncbi.nlm.nih.gov/29787309/)]
53. Brennan TA, Localio RJ, Laird NL. Reliability and validity of judgments concerning adverse events suffered by hospitalized patients. *Med Care* 1989 Dec;27(12):1148-1158. [doi: [10.1097/00005650-198912000-00006](https://doi.org/10.1097/00005650-198912000-00006)] [Medline: [2593729](https://pubmed.ncbi.nlm.nih.gov/2593729/)]
54. Haley RW, Schaberg DR, McClish DK, Quade D, Crossley KB, Culver DH, et al. The accuracy of retrospective chart review in measuring nosocomial infection rates. Results of validation studies in pilot hospitals. *Am J Epidemiol* 1980 May;111(5):516-533. [doi: [10.1093/oxfordjournals.aje.a112931](https://doi.org/10.1093/oxfordjournals.aje.a112931)] [Medline: [7377196](https://pubmed.ncbi.nlm.nih.gov/7377196/)]

Abbreviations

- AUC:** area under the curve
- AUC-ROC:** area under the receiver operating characteristic curve
- AU-PRC:** area under the precision-recall curve
- BERT:** bidirectional encoder representations from transformers
- BOW:** bag-of-words
- DAD:** discharge abstract database
- EMR:** electronic medical record
- HAN:** hierarchical attention network

HAPIs: hospital-acquired pressure injuries
ICD: International Classification of Diseases
ICU: intensive care unit
ML: machine learning
NLP: natural language processing
NPV: negative predictive value
PI: pressure injury
PPV: positive predictive value
RCT: randomized controlled trial
RF: random forest
SCM: Sunrise Clinical Manager
TF-IDF: term frequency-inverse document frequency
XGBoost: extreme gradient boosting

Edited by K El Emam, B Malin; submitted 24.07.22; peer-reviewed by B Ru, T Zhang; comments to author 05.10.22; revised version received 01.01.23; accepted 15.01.23; published 08.03.23.

Please cite as:

*Nurmambetova E, Pan J, Zhang Z, Wu G, Lee S, Southern DA, Martin EA, Ho C, Xu Y, Eastwood CA
Developing an Inpatient Electronic Medical Record Phenotype for Hospital-Acquired Pressure Injuries: Case Study Using Natural Language Processing Models
JMIR AI 2023;2:e41264
URL: <https://ai.jmir.org/2023/1/e41264>
doi: [10.2196/41264](https://doi.org/10.2196/41264)
PMID:*

©Elvira Nurmambetova, Jie Pan, Zilong Zhang, Guosong Wu, Seungwon Lee, Danielle A Southern, Elliot A Martin, Chester Ho, Yuan Xu, Cathy A Eastwood. Originally published in JMIR AI (<https://ai.jmir.org>), 08.03.2023. This is an open-access article distributed under the terms of the Creative Commons Attribution License (<https://creativecommons.org/licenses/by/4.0/>), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR AI, is properly cited. The complete bibliographic information, a link to the original publication on <https://www.ai.jmir.org/>, as well as this copyright and license information must be included.

Original Paper

Automated Identification of Aspirin-Exacerbated Respiratory Disease Using Natural Language Processing and Machine Learning: Algorithm Development and Evaluation Study

Thanai Pongdee¹, MD; Nicholas B Larson², MS, PhD; Rohit Divekar¹, MBBS, PhD; Suzette J Bielinski³, PhD; Hongfang Liu⁴, PhD; Sungrim Moon⁴, PhD

¹Division of Allergic Diseases, Mayo Clinic, Rochester, MN, United States

²Division of Clinical Trials and Biostatistics, Department of Quantitative Health Sciences, Mayo Clinic, Rochester, MN, United States

³Division of Epidemiology, Department of Quantitative Health Sciences, Mayo Clinic, Rochester, MN, United States

⁴Department of Artificial Intelligence and Informatics, Mayo Clinic, Rochester, MN, United States

Corresponding Author:

Thanai Pongdee, MD

Division of Allergic Diseases

Mayo Clinic

200 First Street SW

Rochester, MN, 55905

United States

Phone: 1 5072843783

Email: pongdee.thanai@mayo.edu

Abstract

Background: Aspirin-exacerbated respiratory disease (AERD) is an acquired inflammatory condition characterized by the presence of asthma, chronic rhinosinusitis with nasal polyposis, and respiratory hypersensitivity reactions on ingestion of aspirin or other nonsteroidal anti-inflammatory drugs (NSAIDs). Despite AERD having a classic constellation of symptoms, the diagnosis is often overlooked, with an average of greater than 10 years between the onset of symptoms and diagnosis of AERD. Without a diagnosis, individuals will lack opportunities to receive effective treatments, such as aspirin desensitization or biologic medications.

Objective: Our aim was to develop a combined algorithm that integrates both natural language processing (NLP) and machine learning (ML) techniques to identify patients with AERD from an electronic health record (EHR).

Methods: A rule-based decision tree algorithm incorporating NLP-based features was developed using clinical documents from the EHR at Mayo Clinic. From clinical notes, using NLP techniques, 7 features were extracted that included the following: AERD, asthma, NSAID allergy, nasal polyps, chronic sinusitis, elevated urine leukotriene E4 level, and documented no-NSAID allergy. MedTagger was used to extract these 7 features from the unstructured clinical text given a set of keywords and patterns based on the chart review of 2 allergy and immunology experts for AERD. The status of each extracted feature was quantified by assigning the frequency of its occurrence in clinical documents per subject. We optimized the decision tree classifier's hyperparameters cutoff threshold on the training set to determine the representative feature combination to discriminate AERD. We then evaluated the resulting model on the test set.

Results: The AERD algorithm, which combines NLP and ML techniques, achieved an area under the receiver operating characteristic curve score, sensitivity, and specificity of 0.86 (95% CI 0.78-0.94), 80.00 (95% CI 70.82-87.33), and 88.00 (95% CI 79.98-93.64) for the test set, respectively.

Conclusions: We developed a promising AERD algorithm that needs further refinement to improve AERD diagnosis. Continued development of NLP and ML technologies has the potential to reduce diagnostic delays for AERD and improve the health of our patients.

(JMIR AI 2023;2:e44191) doi:[10.2196/44191](https://doi.org/10.2196/44191)

KEYWORDS

aspirin exacerbated respiratory disease; natural language processing; electronic health record; identification; machine learning; aspirin; asthma; respiratory illness; artificial intelligence; natural language processing algorithm

Introduction

Aspirin-exacerbated respiratory disease (AERD) is an acquired inflammatory condition characterized by the presence of asthma, chronic rhinosinusitis with nasal polyposis, and respiratory hypersensitivity reactions on ingestion of aspirin or other nonsteroidal anti-inflammatory drugs (NSAIDs) [1]. These reactions typically involve the upper and lower airways and may include nasal congestion, sneezing, rhinorrhea, cough, and wheezing [1]. The prevalence of AERD is approximately 0.3%-0.9% in the general population, but the actual prevalence is unknown in practice, as AERD has no unique International Classification of Diseases, Ninth Revision (ICD-9) or ICD-10 codes [2,3]. In the general population, the mean age of onset of AERD is approximately 30 years [2,4], and the prevalence of AERD is estimated to be 7%-15% in individuals with asthma and 10%-16% in individuals with chronic rhinosinusitis with nasal polyposis [5]. Individuals with AERD have significant symptom burden and morbidity, including severe and recalcitrant sinus disease, high rates of polyp recurrence and revision surgery, and higher asthma exacerbation and hospitalization rates [1]. Despite AERD having a classic constellation of symptoms, the diagnosis is often overlooked, with an average of greater than 10 years between the onset of symptoms and diagnosis of AERD [6]. Without a diagnosis, individuals will lack opportunities to receive effective treatments, such as aspirin desensitization or biologic medications [5,7].

One opportunity to improve diagnostic delays with AERD involves leveraging the immense volume of clinical data available in electronic health records (EHRs). By leveraging natural language processing (NLP) and machine learning (ML), analyses of medical concepts from unstructured clinical documents may aid in early detection of AERD [8]. In this study, we developed a combined algorithm of NLP with ML to identify individuals with AERD.

Methods

Ethical Considerations

This study was approved by the Mayo Clinic institutional review board as exempted from ethics approval in accordance with the ethical standards of the responsible committee on human experimentation and the Helsinki Declaration of 1975, as revised in 2000.

Procedure

Patients who were evaluated within the Allergy and Immunology divisions at Mayo Clinic from January 2001 to March 2022 and met diagnostic criteria for AERD based on accepted guidelines [1] were retrospectively identified by chart review. In total, 200 patients with AERD and 200 patients without AERD were identified. Of these patients, we randomly selected 100 patients

with AERD and 100 without AERD to serve as the training set, and the remaining were used for the test set.

A rule-based decision tree algorithm incorporating NLP-based features was developed to identify patients with AERD using clinical documents from the EHR at Mayo Clinic. From clinical notes, 7 features were extracted using NLP techniques based on common characteristics of AERD [1]. These features included the following: prior AERD diagnosis, asthma, NSAID allergy, nasal polyps, chronic sinusitis, elevated urine leukotriene E4 level, and documented no-NSAID allergy. "Prior AERD diagnosis" was defined as whether the patient had a diagnosis of AERD before or had suspicion of a high chance of AERD by the physician. For "asthma," "nasal polyps," and "chronic sinusitis," the patient needed to have a diagnosis confirmation by the physician in the clinical documents. "Elevated urine leukotriene E4 level" indicated if the patient had any record in lab results of a urine leukotriene E4 level greater than 104 pg/mg creatinine. "NSAID allergy" was defined as a patient having had a respiratory reaction to an NSAID. Meanwhile, "documented no-NSAID allergy" indicated that a health care provider recorded "unconfirmed or no specific history of NSAID allergy up to date" in the clinical documents. Given the successful use cases of MedTagger [9] to identify disease in different clinical domains [10,11], we used MedTagger to extract these features with the given set of keywords (including typos, abbreviations, and acronyms) and patterns based on the chart review of 2 allergy and immunology experts for AERD. If the extracted features were located in particular note sections (ie, "History of Present Illness," "Allergies," "Past Medical/Surgical History," "Impression/Report/Plan," "Diagnosis," "Principal Diagnosis," "Secondary Diagnoses," and "Post Procedure Diagnosis"), they were considered valid AERD features. We collected each feature in all clinical documents per patient in the past 5 years from the last clinic visit because clinical characteristics of AERD can evolve over time (ie, development of NSAID allergy).

We counted the number of times each extracted feature appeared in the clinical documents for each patient and used this count as the numerical representation of each feature. To identify the most practical combination of features for discriminating between different presentations of AERD, we optimized the hyperparameters of the classification and regression tree (CART) decision tree classifier with the identified features on the training set using sklearn [12,13]. We performed hyperparameter tuning on 5 different parameters with 1 model setting, as follows: (1) criterion, with options of gini or entropy; (2) maximum depth, ranging from 1 to 10 with an interval of 1; (3) minimum samples split, ranging from 2 to 10 with an interval of 2; (4) minimum samples leaf, ranging from 1 to 10 with an interval of 1; (5) maximum features, ranging from 1 to 7 with an interval of 1; and (6) a fixed random number generation seed was used to ensure reproducibility. Furthermore, to achieve the highest area under the receiver operating characteristics curve (AUC) score, these hyperparameters were tuned for two types of feature sets:

(1) quantitatively represented as numerical values per patient and (2) binary, where “1” denotes the presence and “0” denotes the absence or missing status of each extracted feature per patient. We constructed a decision tree using the best feature set with optimized hyperparameters and then calculated the AUC scores for a range of cutoff thresholds from 0.1 to 1.0 in intervals of 0.1 to determine the optimal cutoff threshold based on a given training set. The resulting tree with the optimized parameters and cutoff threshold converted into sequential rule sets to evaluate the performance in the test set.

Results

In our cohort, the mean age of the 400 patients was 55.5 years, and 54% (216/400) were female. [Table 1](#) displays the descriptive statistics for each feature, comparing the presence or absence of the feature in the training and test sets. Based on the training set, we obtained the sequential rule sets through the optimized decision tree (with criterion as gini, maximum depth as 7, minimum samples leaf as 7, minimum samples split as 2, maximum features as 3, random state as 20, and best cutoff threshold as 0.6 for parameter settings) using the numerical represented feature set in [Table 2](#). The sequential rules listed in

[Table 2](#) describe several clinical factors that include diagnosis of AERD (referred to as AERD), diagnosis of allergy to an NSAID (referred to as NSAID allergy), diagnosis of chronic sinusitis, documented history of tolerance to an NSAID (referred to as non-NSAID allergy), and a prior abnormally elevated urine leukotriene E4 level (referred to as LAB).

In [Table 2](#), it was observed that the derived sequential rule, ranging from 1 to 9, captured 28% (56/200) of the cases in the test set. However, a significant portion of the test set (112/200, 56%) was not identified according to the original intended sequential rule but rather by a different sequence rule. For example, rule 6 failed to capture 73 cases, whereas rule 9—which is less strict than rule 6—captured 59 of those 73 cases that were supposed to belong to rule 6. Similarly, rule 3 captured 15 cases of the remaining 18 cases that should have been identified by rule 1. Therefore, the overall accuracy was 0.84.

The AERD algorithm achieved an AUC score of 0.92 (95% CI 0.93-1.00) and 0.86 (95% CI 0.78-0.94) for the training and test sets ([Figure 1](#) and [Figure 2](#)), respectively. The optimal cutoff point was 0.6 on the training set ([Figure 1](#)). Additional performances are presented in [Table 3](#).

Table 1. Descriptive statistics of aspirin-exacerbated respiratory disease (AERD) features, describing its presence as 1 and absence as 0 (N=200).

AERD Feature	Train, n (%)	Test, n (%)
AERD	103 (52)	60 (30)
Asthma	192 (96)	82 (41)
NSAID ^a allergy	98 (49)	121 (61)
Nasal polyps	175 (88)	192 (96)
Chronic sinusitis	182 (91)	180 (90)
LAB ^b	93 (47)	179 (90)
Documented no-NSAID allergy	70 (35)	101 (51)

^aNSAID: nonsteroidal anti-inflammatory drug.

^bLAB refers to the elevated urine leukotriene E4 level.

Table 2. Derived sequential rules for aspirin-exacerbated respiratory disease (AERD) algorithm and the resulting performance in the test set.

Rule	Sequential rules	AERD	Case, n	Correct, n	Error, n	Confidence (%) ^a
1	AERD \leq 3.5, NSAID allergy ^b \leq 2.5, Chronic Sinusitis ^c \leq 6.5, and then documented non-NSAID allergy \leq 0.5	No	30	12	0	40
2	AERD \leq 3.5, NSAID allergy \leq 2.5, Chronic Sinusitis \leq 6.5, and then documented non-NSAID allergy $>$ 0.5	No	9	0	0	0
3	AERD \leq 3.5, NSAID allergy \leq 2.5, Chronic Sinusitis $>$ 6.5, and then documented non-NSAID allergy \leq 0.5	No	43	40	0	93
4	AERD \leq 3.5, NSAID allergy \leq 2.5, Chronic Sinusitis $>$ 6.5, and then documented non-NSAID allergy $>$ 0.5	No	4	3	0	75
5	AERD \leq 3.5, NSAID allergy $>$ 2.5, and then Chronic Sinusitis \leq 9.0	Yes	10	0	0	0
6	AERD \leq 3.5, NSAID allergy $>$ 2.5, and then Chronic Sinusitis $>$ 9.0	Yes	74	1	0	1
7	AERD $>$ 3.5, NSAID allergy \leq 1.5, and then LAB ^d \leq 0.5	Yes	0	0	0	0
8	AERD $>$ 3.5, NSAID allergy \leq 1.5, and then LAB $>$ 0.5	No	2	0	0	0
9	AERD $>$ 3.5, NSAID allergy $>$ 1.5	Yes	2	0	0	0
10	Others	Yes	16	0	0	0
		No	10	0	0	0
N/A ^e	The cases were not identified according to the original intended sequential rule; instead, a different sequence rule was used.	Yes	99	79	20	80
		No	45	33	12	73

^aConfidence = the numbers of correct cases divided by numbers of real cases in the test set multiplied by 100 for the particular rule from 1 to 9.

^bNSAID allergy refers to diagnosis of allergy to a nonsteroidal anti-inflammatory drug (NSAID).

^cChronic sinusitis refers to diagnosis of chronic sinusitis.

^dLAB refers to a prior abnormally elevated urine leukotriene E4 level.

^eN/A: not applicable.

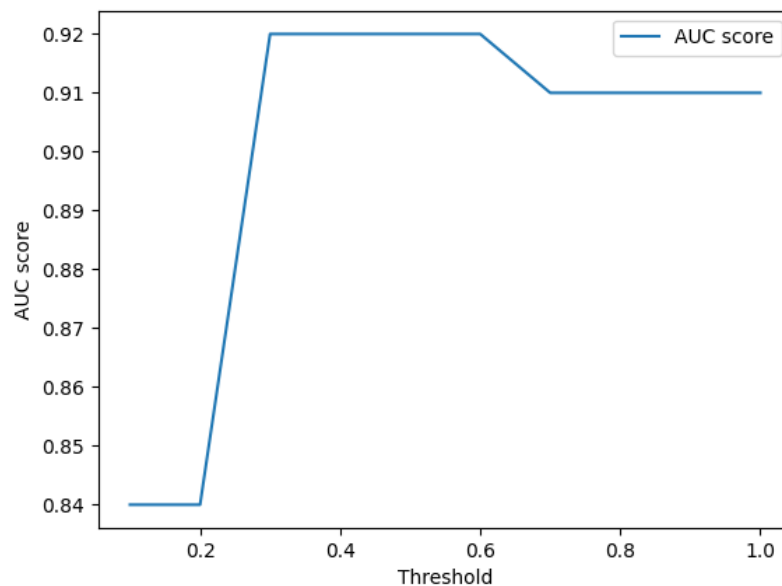
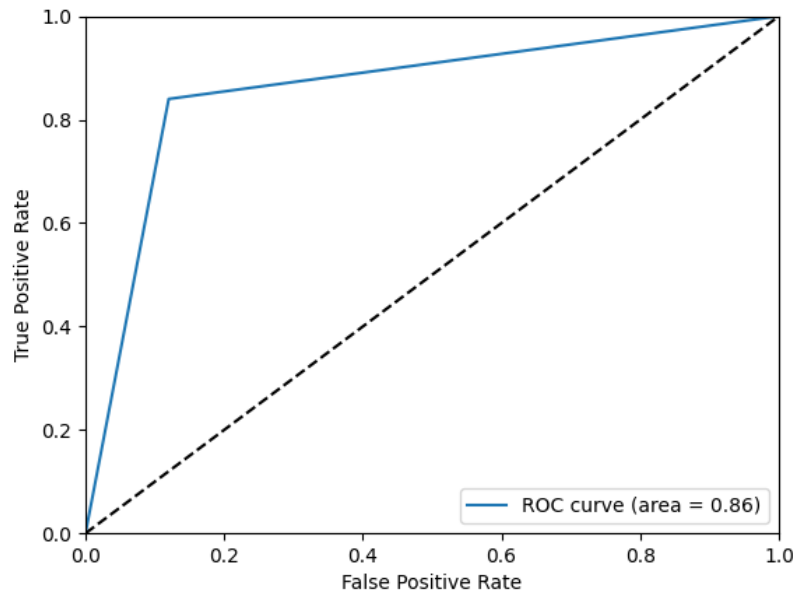
Figure 1. Area under the receiver operating characteristic curve (AUC) scores at different threshold values on the training set.

Figure 2. Receiver operating characteristic (ROC) on the test set.**Table 3.** Performance of the rule-based aspirin-exacerbated respiratory disease (AERD) algorithm.

Data set	Sensitivity (%; 95% CI)	Specificity (%; 95% CI)	Positive predictive value (%; 95% CI)	Negative predictive value (%; 95% CI)	Accuracy (%; 95% CI)
Train	88.00 (79.98-93.64)	97.00 (91.48-99.38)	96.70 (90.67-99.31)	88.99 (81.56-94.18)	92.50 (87.93-95.74)
Test	80.00 (70.82-87.33)	88.00 (79.98-93.64)	86.96 (78.32-93.07)	81.48 (72.86-88.31)	84.00 (78.17-88.79)

Discussion

Principal Findings

In our study, we demonstrated that an algorithm, which combines NLP and ML techniques, can identify patients with AERD with a positive predictive value of approximately 86.96 and a negative predictive value of 81.48. Our results are comparable to prior work [3] on automated diagnosis of AERD from EHR data using structured query language statements for data analysis and resulting in positive predictive values ranging from 78.4 to 88.7, depending on the cohort being analyzed.

Prior diagnosis of AERD presents the highest impacted feature (ie, a majority of sequential rules contain prior diagnosis of AERD feature) to detect diagnosis of AERD. In the training and test sets, 85% (85/100) and 91% (91/100) of patients with AERD had a prior diagnosis of AERD, respectively. We also extracted new clinical factors associated with AERD (“elevated urine leukotriene E4 level” and “alcohol intolerance”) that were not previously studied. Furthermore, the “elevated urine leukotriene E4 level” feature may need to be considered as a new meaningful feature associated with AERD because the presence of the term “AERD” with an “elevated leukotriene E4 level” was a common feature of rule sets 7 and 8. Most patients with AERD in the test set were accurately identified by having had an AERD diagnosis and a documented NSAID allergy (Table 2). Lastly, diagnosis of nasal polyps was not used to construct the optimal decision tree, which may indicate that it may be an insignificant feature to distinguish patients with AERD from possible AERD candidates.

The test set included 32 errors from 200 patients, which upon review, were due to either unidentified rule sets for patients with AERD (n=11) or missing and incorrect feature extraction because of unseen keywords or patterns for features (n=9) primarily. For example, the sentence “Patient took an aspirin approximately ten years ago for headache and developed a sensation of pressure in his nose and sinuses” is an unseen pattern for prior AERD features. Based on the expression, “a sensation of pressure in his nose and sinuses,” the sentence should be a prior AERD feature; however, AERD algorithm categorized it as absence of an AERD feature because this pattern was not available in the training phrase. A total of 6 patients had necessary feature information beyond the past 5 years of clinical documents from the last visit day; 6 patients had necessary information belonging to an unknown note section in the training set for feature extraction. When examining the specific rules, rule sets 2-3 resulted in very few errors (Table 2). In contrast, the absence of terms explicitly documenting the absence of NSAID allergy and lack of references to an elevated leukotriene E4 level resulted in more errors in the AERD algorithm.

Diagnosing and confirming AERD may be a prolonged process, as the associated clinical features may present at different times in a variety of time sequences. As a result, there is no solid ICD code (structured data) to represent AERD, and AERD-associated clinical characteristics are often undocumented in clinical texts (unstructured data) in the EHR. This lack of information regarding AERD results in the low quality of data sources and potential bias for ML models [14]. Additional efforts (eg, standardizing routine exams for AERD) are necessary to fill these missing information gaps in practice.

This AERD algorithm has limitations in deploying to detect patients with confirmed AERD in a practical setting without further refinement. We focused on identifying feature selections in the limited parameter tuning using a balanced data set (N=200 for patients with AERD and N=200 for patients without AERD), which was not a real-world situation. We used the minimum sample size due to the nature of AERD, which has a low prevalence. The rule-based algorithm is used because the limited sample and feature set provide high interpretability and accuracy at downstream tasks rather than neural network MLs, which require a large training data set. However, this algorithm provides a valuable contribution to capturing potential patients with AERD in the setting of a large health system EHR because the prevalence of patients with AERD is low in clinical settings. To follow up, we plan to rank features with diverse identified

feature sets and parameter tuning for the decision tree model within a large cohort. We will investigate our new feature in the EHR, which is information about urine leukotriene E4 levels in the extensive feature selections, and we will explore additional features for AERD (eg, alcohol sensitivity, anosmia, and prior sinus surgeries).

Conclusions

We developed an AERD algorithm, which combines NLP and ML techniques, to enhance AERD diagnosis in practice. On top of prior work [3], we used NLP with a potential feature—urine leukotriene E4 levels from EHR—which have been shown to aid in AERD diagnosis [15]. Leveraging NLP and ML techniques in practice has the potential to reduce diagnostic delays for AERD and improve the health of patients.

Acknowledgments

All authors had substantial contributions to the design, acquisition, analysis, and interpretation of data for this study. In addition, all authors contributed to the drafts, revisions, and approval of the final version to be published.

This work was supported by grants from the National Heart, Lung and Blood Institute (R01 HL136659) and the resources of the Rochester Epidemiology Project (REP) medical records-linkage system, which is supported by the National Institute on Aging (NIA; AG 058738), by the Mayo Clinic Research Committee, and by fees paid annually by REP users. The funding sources played no role in the design, conduct, or reporting of this study. The content of this paper is solely the responsibility of the authors and does not represent the official views of the National Institutes of Health (NIH) or the Mayo Clinic.

Conflicts of Interest

HL is the Associate Editor of *JMIR AI*. The other authors declare that they have no conflicts of interest.

References

1. Haque R, White AA, Jackson DJ, Hopkins C. Clinical evaluation and diagnosis of aspirin-exacerbated respiratory disease. *J Allergy Clin Immunol* 2021 Aug;148(2):283-291. [doi: [10.1016/j.jaci.2021.06.018](https://doi.org/10.1016/j.jaci.2021.06.018)] [Medline: [34364538](https://pubmed.ncbi.nlm.nih.gov/34364538/)]
2. Szczeklik A, Nizankowska E, Duplaga M. Natural history of aspirin-induced asthma. AIANE Investigators. European Network on Aspirin-Induced Asthma. *Eur Respir J* 2000 Sep;16(3):432-436 [FREE Full text] [doi: [10.1034/j.1399-3003.2000.016003432.x](https://doi.org/10.1034/j.1399-3003.2000.016003432.x)] [Medline: [11028656](https://pubmed.ncbi.nlm.nih.gov/11028656/)]
3. Cahill KN, Johns CB, Cui J, Wickner P, Bates DW, Laidlaw TM, et al. Automated identification of an aspirin-exacerbated respiratory disease cohort. *J Allergy Clin Immunol* 2017 Mar;139(3):819-825.e6 [FREE Full text] [doi: [10.1016/j.jaci.2016.05.048](https://doi.org/10.1016/j.jaci.2016.05.048)] [Medline: [27567328](https://pubmed.ncbi.nlm.nih.gov/27567328/)]
4. Berges-Gimeno MP, Simon RA, Stevenson DD. The natural history and clinical characteristics of aspirin-exacerbated respiratory disease. *Ann Allergy Asthma Immunol* 2002 Nov;89(5):474-478. [doi: [10.1016/s1081-1206\(10\)62084-4](https://doi.org/10.1016/s1081-1206(10)62084-4)]
5. Stevens WW, Jerschow E, Baptist AP, Borish L, Bosso JV, Buchheit KM, et al. *J Allergy Clin Immunol* 2021 Mar;147(3):827-844 [FREE Full text] [doi: [10.1016/j.jaci.2020.10.043](https://doi.org/10.1016/j.jaci.2020.10.043)] [Medline: [33307116](https://pubmed.ncbi.nlm.nih.gov/33307116/)]
6. Lee-Sarwar K, Johns C, Laidlaw TM, Cahill KN. Tolerance of daily low-dose aspirin does not preclude aspirin-exacerbated respiratory disease. *J Allergy Clin Immunol Pract* 2015 May;3(3):449-451 [FREE Full text] [doi: [10.1016/j.jaip.2015.01.007](https://doi.org/10.1016/j.jaip.2015.01.007)] [Medline: [25634222](https://pubmed.ncbi.nlm.nih.gov/25634222/)]
7. Buchheit KM, Laidlaw TM, Levy JM. Immunology-based recommendations for available and upcoming biologics in aspirin-exacerbated respiratory disease. *J Allergy Clin Immunol* 2021 Aug;148(2):348-350 [FREE Full text] [doi: [10.1016/j.jaci.2021.06.019](https://doi.org/10.1016/j.jaci.2021.06.019)] [Medline: [34174296](https://pubmed.ncbi.nlm.nih.gov/34174296/)]
8. Khoury P, Srinivasan R, Kakumanu S, Ochoa S, Keswani A, Sparks R, et al. A framework for augmented intelligence in allergy and immunology practice and research—a work group report of the AAAAI Health Informatics, Technology, and Education Committee. *J Allergy Clin Immunol Pract* 2022 May;10(5):1178-1188. [doi: [10.1016/j.jaip.2022.01.047](https://doi.org/10.1016/j.jaip.2022.01.047)] [Medline: [35300959](https://pubmed.ncbi.nlm.nih.gov/35300959/)]
9. Wen A, Fu S, Moon S, El Wazir M, Rosenbaum A, Kaggal VC, et al. Desiderata for delivering NLP to accelerate healthcare AI advancement and a Mayo Clinic NLP-as-a-service implementation. *NPJ Digit Med* 2019 Dec 17;2(1):130 [FREE Full text] [doi: [10.1038/s41746-019-0208-8](https://doi.org/10.1038/s41746-019-0208-8)] [Medline: [31872069](https://pubmed.ncbi.nlm.nih.gov/31872069/)]

10. Afzal N, Mallipeddi VP, Sohn S, Liu H, Chaudhry R, Scott CG, et al. Natural language processing of clinical notes for identification of critical limb ischemia. *Int J Med Inform* 2018 Mar;111:83-89 [FREE Full text] [doi: [10.1016/j.ijmedinf.2017.12.024](https://doi.org/10.1016/j.ijmedinf.2017.12.024)] [Medline: [29425639](https://pubmed.ncbi.nlm.nih.gov/29425639/)]
11. Moon S, Liu S, Scott CG, Samudrala S, Abidian MM, Geske JB, et al. Automated extraction of sudden cardiac death risk factors in hypertrophic cardiomyopathy patients by natural language processing. *Int J Med Inform* 2019 Aug;128:32-38 [FREE Full text] [doi: [10.1016/j.ijmedinf.2019.05.008](https://doi.org/10.1016/j.ijmedinf.2019.05.008)] [Medline: [31160009](https://pubmed.ncbi.nlm.nih.gov/31160009/)]
12. Decision trees: scikit-learn. URL: <https://scikit-learn.org/stable/modules/generated/sklearn.tree.DecisionTreeClassifier.html> [accessed 2023-01-05]
13. Breiman L, Friedman J, Olshen R, Stone C. *Classification and Regression Trees* (1st Edition). Oxfordshire, UK: Routledge; 1984.
14. Moon S, Carlson LA, Moser ED, Agnikula Kshatriya BS, Smith CY, Rocca WA, et al. Identifying information gaps in electronic health records by using natural language processing: gynecologic surgery history identification. *J Med Internet Res* 2022 Jan 28;24(1):e29015 [FREE Full text] [doi: [10.2196/29015](https://doi.org/10.2196/29015)] [Medline: [35089141](https://pubmed.ncbi.nlm.nih.gov/35089141/)]
15. Divekar R, Hagan J, Rank M, Park M, Volcheck G, O'Brien E, et al. Diagnostic utility of urinary LTE4 in asthma, allergic rhinitis, chronic rhinosinusitis, nasal polyps, and aspirin sensitivity. *J Allergy Clin Immunol Pract* 2016 Jul;4(4):665-670 [FREE Full text] [doi: [10.1016/j.jaip.2016.03.004](https://doi.org/10.1016/j.jaip.2016.03.004)] [Medline: [27080204](https://pubmed.ncbi.nlm.nih.gov/27080204/)]

Abbreviations

AERD: aspirin-exacerbated respiratory disease
AUC: area under the receiver operating characteristic curve
EHR: electronic health record
ICD: International Classification of Diseases
ML: machine learning
NLP: natural language processing
NSAID: nonsteroidal anti-inflammatory drug

Edited by B Malin, K El Emam; submitted 09.11.22; peer-reviewed by GK Ramachandran, L Tong; comments to author 26.11.22; revised version received 19.01.23; accepted 22.05.23; published 12.06.23.

Please cite as:

Pongdee T, Larson NB, Divekar R, Bielinski SJ, Liu H, Moon S

Automated Identification of Aspirin-Exacerbated Respiratory Disease Using Natural Language Processing and Machine Learning: Algorithm Development and Evaluation Study

JMIR AI 2023;2:e44191

URL: <https://ai.jmir.org/2023/1/e44191>

doi: [10.2196/44191](https://doi.org/10.2196/44191)

©Thanai Pongdee, Nicholas B Larson, Rohit Divekar, Suzette J Bielinski, Hongfang Liu, Sungrim Moon. Originally published in JMIR AI (<https://ai.jmir.org>), 12.06.2023. This is an open-access article distributed under the terms of the Creative Commons Attribution License (<https://creativecommons.org/licenses/by/4.0/>), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR AI, is properly cited. The complete bibliographic information, a link to the original publication on <https://www.ai.jmir.org/>, as well as this copyright and license information must be included.

Original Paper

Extractive Clinical Question-Answering With Multianswer and Multifocus Questions: Data Set Development and Evaluation Study

Sungrim Moon¹, PhD; Huan He¹, PhD; Heling Jia¹, MD; Hongfang Liu¹, PhD; Jungwei Wilfred Fan¹, PhD

Department of Artificial Intelligence & Informatics, Mayo Clinic, Rochester, MN, United States

Corresponding Author:

Jungwei Wilfred Fan, PhD

Department of Artificial Intelligence & Informatics

Mayo Clinic

200 1st Street SW

RO_HA_07_741B-I

Rochester, MN, 55905

United States

Phone: 1 507 538 1191

Email: Fan.Jung-wei@mayo.edu

Abstract

Background: Extractive question-answering (EQA) is a useful natural language processing (NLP) application for answering patient-specific questions by locating answers in their clinical notes. Realistic clinical EQA can yield multiple answers to a single question and multiple focus points in 1 question, which are lacking in existing data sets for the development of artificial intelligence solutions.

Objective: This study aimed to create a data set for developing and evaluating clinical EQA systems that can handle natural multianswer and multifocus questions.

Methods: We leveraged the annotated relations from the 2018 National NLP Clinical Challenges corpus to generate an EQA data set. Specifically, the 1-to-N, M-to-1, and M-to-N drug-reason relations were included to form the multianswer and multifocus question-answering entries, which represent more complex and natural challenges in addition to the basic 1-drug-1-reason cases. A baseline solution was developed and tested on the data set.

Results: The derived RxWhyQA data set contains 96,939 QA entries. Among the answerable questions, 25% of them require multiple answers, and 2% of them ask about multiple drugs within 1 question. Frequent cues were observed around the answers in the text, and 90% of the *drug* and *reason* terms occurred within the same or an adjacent sentence. The baseline EQA solution achieved a best F_1 -score of 0.72 on the entire data set, and on specific subsets, it was 0.93 for the unanswerable questions, 0.48 for single-drug questions versus 0.60 for multidrug questions, and 0.54 for the single-answer questions versus 0.43 for multianswer questions.

Conclusions: The RxWhyQA data set can be used to train and evaluate systems that need to handle multianswer and multifocus questions. Specifically, multianswer EQA appears to be challenging and therefore warrants more investment in research. We created and shared a clinical EQA data set with multianswer and multifocus questions that would channel future research efforts toward more realistic scenarios.

(JMIR AI 2023;2:e41818) doi:[10.2196/41818](https://doi.org/10.2196/41818)

KEYWORDS

question-answering; information extraction; dataset; data set; artificial intelligence; natural language processing

Introduction

Background

The thought process involved in clinical reasoning and decision-making can be naturally framed into a series of questions and answers [1,2]. Achieving human-like question-answering (QA) capability is highly regarded in artificial intelligence (AI). Medical QA research has garnered terrific momentum over the past decade, and a new generation of AI scientists is undergoing a state-of-the-art update at a daunting pace almost every month (if not every week). One of the very sought-after applications is to find the answer within a given document, or extractive QA (EQA), which enables patient-specific QA based on the information provided in the clinical text [3]. As an essential component in most AI engineering undertakings, EQA training data determine not only the likelihood of success in terms of annotation quality but also the fidelity of representing the target scenario.

Along with other issues observed in existing medical EQA corpora [4], the mainstream annotation approach knowingly simplifies the task into a “one answer per document” scheme. Although the simplification makes development and evaluation easier for promoting initial growth of the field, it is unrealistic because EQA can naturally have multiple qualified answers (or answer components) within 1 document, and often all of them must be captured to sufficiently answer a question [5]. Moreover, a question can naturally involve multiple focus points such as “Why A, B, and C...” rather than requiring the user to ask 1 question for each point. To address this gap, we created an EQA data set that involves realistic, multianswer and multifocus cases by converting the concept-relation annotations from an existing clinical natural language processing (NLP) challenge data set. Our generated RxWhyQA data set includes a total of 96,939 QA entries, where 25% of the answerable questions require the identification of multiple answers and 2% of them ask about multiple drugs within 1 question. We also developed a baseline solution for multianswer QA and tested it on the RxWhyQA.

The novelty of this study is reframing the original relation identification task into an EQA task, which simplifies the conventional 2-step approach of named entity recognition and relation classification into 1-step information extraction guided by natural language questions. Our primary contribution is the RxWhyQA as a resource that offers realistic constructs to facilitate NLP research in this underexplored area. To our knowledge, there has not been any EQA data set that contains multianswer and multifocus questions based on clinical notes.

Related Work

QA is a versatile task that can subsume diverse NLP tasks when properly represented [6]. More than a decade of research has focused on the EQA task in NLP [7]. As the name implies, EQA can be viewed as question-guided information extraction from a given text. Unlike conventional approaches that require the identification of the involved entities as one task followed by determination of the target relation between the entities as the other task, EQA consolidates these steps into a smooth one-shot task where the user asks a natural language question for the

system to understand the focus point, identify relevant cues in the text, and locate the answer that satisfies the relation of interest. Although EQA demands higher machine intelligence, it is efficient in terms of the data schema for modeling and the human-computer interaction for users.

The Stanford Question Answering Dataset (SQuAD) [8] established a widely adopted framework for EQA, and in the later version (version 2.0) [9], the task also requires a system to refrain from answering when no suitable answer is present in the text. In the clinical domain, corpora have been developed for EQA based on electronic health records (EHRs). In the study by Raghavan et al [10], medical students were presented with structured and unstructured EHR information about each patient to generate realistic questions for a hypothetical office encounter. Using the BioASQ data set based on biomedical literature, Yoon et al [5] proposed a sequence tagging approach to handling multianswer EQA. In the consumer health domain, Zhu et al [11] developed a Multiple Answer Spans Healthcare Question Answering (ie, MASH-QA) data set specifically involving multiple answers of nonconsecutive spans in the target text. As a non-English example, Ju et al [12] developed a Conditional Multiple-span Chinese Question Answering data set from a web-based QA forum. Pampari et al [13] developed the emrQA, a large clinical EQA corpus generated through template-based semantic extraction from the Informatics for Integrating Biology & the Bedside NLP challenge data sets. We took a similar approach as the emrQA but additionally included multianswer and multifocus questions that better reflect natural clinical EQA scenarios.

Methods

Generating the QA Annotations From a Relation Identification Challenge

Our source data were based on the annotations originally created for the National NLP Clinical Challenges (n2c2) corpus of 2018, which aimed to identify adverse drug events by extracting various drug-related concepts and classifying their relations in the clinical text [14]. Their final gold standard included 83,869 concepts and 59,810 relations in 505 discharge summaries. In this study, we focused on generating QA pairs from the subset of drug and reason concepts (ie, mainly about the prescribing justification) and the relations between the concepts. Each relation consisted of 2 arguments: a drug concept and a reason concept, as in an example pair such as *drug-reason* (morphine-pain). Accordingly, a question around the drug concept could be derived, such as “Why was morphine prescribed to the patient?” and the reason concept “pain” would be designated as the answer. In the n2c2 corpus, each pair of drug and reason concepts had their text mentions annotated in the corresponding clinical document. The properties make for a good EQA data set where the system is expected to consider the actual contexts surrounding the drug and reason rather than performing a simple lookup. This is especially important for extracting off-label uses because a standard indication knowledge base would not cover those exceptions documented in real-world clinical text.

From the n2c2 annotations on each clinical document, we leveraged several relation types between the drug and reason concepts: 1 drug 0 reason, 1 drug 1 reason, 1 drug N reasons, N drugs 1 reason, or M drugs N reasons. The most straightforward were the 1-drug-1-reason relations (eg, the morphine-pain relation mentioned above), each translated into a 1-to-1 QA entry. The 1-drug-0-reason relations apparently corresponded to the 1-to-0 (unanswerable) QA entries. We preserved the 1-drug-N-reasons relations directly as 1-to-N QAs that require locating multiple answers in the text. For the N-drugs-1-reason and M-drugs-N-reasons relations, we preserved the original multidrug challenge in questions such as, “Why were amlodipine, metoprolol, and isosorbide prescribed to the patient?” The M-drugs-N-reasons relations would also derive multianswer entries such as those derived from the 1-drug-N-reasons relations. In addition to the generated QA entries, we also supplemented paraphrastic questions [15] that may enhance the generalizability of the trained systems.

Quantitative and Qualitative Analysis of the Derived QA Annotations

Along with descriptive statistics of the QA entries and the number of answers per question, we computed the frequencies of the specific drug and reason concept terms (after applying lexical normalization such as lowercase) among the QA entries. The frequencies were meant to offer an intuitive estimate of the abundance of train/test data available for each specific concept or concept pair. We then randomly sampled 100 QA entries for manual review: 50 from those with a single answer and 50 from those with multiple answers. The common patterns informative to QA inference were summarized, offering evidence on what the potential AI solutions could leverage. In addition, we measured the distance (by the number of sentences) between the question and answer concepts. For each specific drug-reason pair, we considered the shortest distance if there were multiple occurrences of either concept. The distance was deemed 0 if the pair occurred within the same sentence. Distance may serve as a surrogate for measuring the challenge to AI systems, where a longer distance implies a more challenging task. In addition, we sampled 100 random drug-reason pairs from each test run (experimental setup described below) to estimate the prevalence of off-label uses in the derived data set. The MEDication-Indication (MEDI) knowledge base (version 2) high-precision subset [16] was first used to screen for on-label uses by exact string match (with normalizing to lowercase), and the remaining drug-reason pairs were reviewed by a domain expert (HJ) to determine off-label uses.

Development of a Baseline Solution

Data Preparation and Model Training

The annotations conform to the SQuAD 2.0 JSON format and can be readily used to train Bidirectional Encoder Representations from Transformers (BERT) [17] for EQA tasks. We randomly partitioned the data set into the train, develop (dev), and test sets by the 5:2:3 ratio, corresponding to 153, 50, and 100 clinical documents, respectively. Random partitioning was carried out 3 times, each executed as a separate run of the experiment for quantifying performance variability. The base language model was ClinicalBERT [18], a domain-customized BERT trained on approximately 2 million clinical documents from the MIMIC-III (version 1.4) database. We fine-tuned ClinicalBERT first on a why-question subset of SQuAD 2.0, followed by fine-tuning on the train set. Training parameters used in the ClinicalBERT fine-tuning were batch_train_size=32, max_seq_length=128, doc_stride=64, learning_rate=3e-5, and epochs=5. The dev set was then used to learn the threshold for determining when the ClinicalBERT model should refrain from providing any answer.

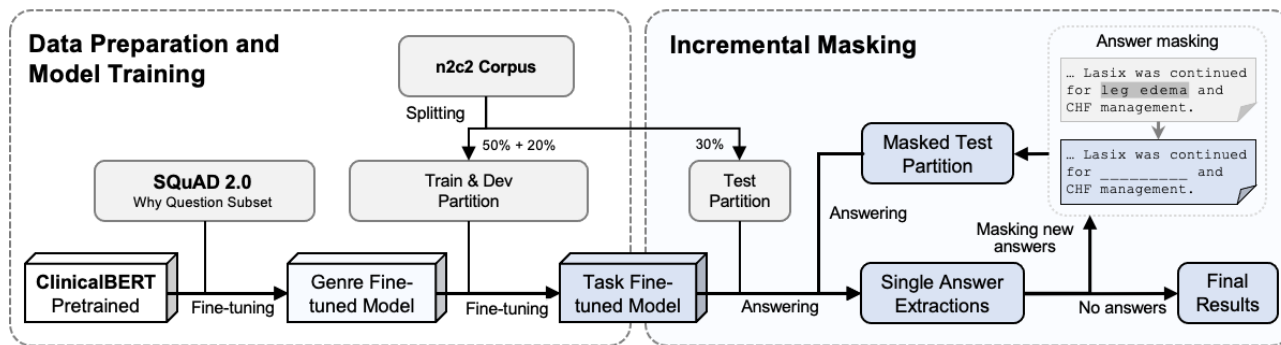
Incremental Masking to Generate Multiple Answers

To force the fine-tuned ClinicalBERT model to continue seeking other suitable answers in each clinical document, we implemented the following process on the test set as a heuristic baseline:

1. Let the EQA model complete its usual single-answer extraction and record the string of the top answer. No further action is needed if the model refrains from answering.
2. Perform a case-insensitive string search using the top answer (from step 1 above) throughout the clinical note from where it was extracted and replace every occurrence into a dummy underscore “_____” string of identical length. This literally generates a new version of the text by masking the original top answer in each question.
3. Run the same EQA model for another round on the entire masked test set again to determine whether the model could identify additional answers elsewhere or started to refrain from answering.

The 3 abovementioned steps were repeated until the model did not generate any new answers on the entire test set. Together, model training and the heuristic multianswer generation process are summarized in [Figure 1](#).

Figure 1. A flowchart of our heuristic approach to constructing a single-answer extractive question-answering model generates multiple answers by incremental masking. The main steps go from left to right. The upper-right “Answer-masking” box illustrates an example of masking where the model’s answer “leg edema” is replaced with a dummy underscore to force the model to look for viable alternative answers elsewhere in the text. BERT: Bidirectional Encoder Representations from Transformers; dev: develop; n2c2: National NLP Clinical Challenges; NLP: natural language processing; SQuAD: Stanford Question Answering Dataset.



Evaluation of the Baseline Solution

After the first round of masking, we began to have more than 1 answer generated by the model for some of the questions. Accordingly, the evaluation program (specifically for the overlap mode) was adapted to accommodate such M-to-N answer comparisons in determining the token-wise proportional match. When anchoring on each gold-standard answer, we selected the model answer with the most overlapping tokens as the best answer in setting the weighted true positive (TP) and false negative (FN); the weighted false positive (FP) was set vice versa by anchoring on each model answer—see equations 1-4 for definitions. On top of these weighted matches between gold-standard and model answers in each question, we tallied them over each entire test set to compute the solution’s precision, recall, and F_1 -score, followed by qualitative error analysis.



Results

Descriptive Statistics of the Derived RxWhyQA Data Set

We leveraged a total of 10,489 relations from the n2c2 adverse drug events NLP challenge and derived the data set, consisting of 96,939 QA entries. Table 1 summarizes the 5 major drug-reason relation categories in the n2c2 corpus, the strategies that we implemented to convert them into QA entries, and their resulting frequencies. Table 2 shows the distribution for the number of answers per question: 75% of the questions have a single answer, while 25% of them require multiple answers. Duplicate answer terms located at different positions of the clinical documents were retained. For example, the procedure “CT” might be mentioned at several places in the text and be recorded as the answer to “Why was the patient prescribed contrast?” We included each such identical term and their different offsets as multiple answers so that the EQA solutions may leverage such nuances. The final data set was formatted into a SQuAD-compatible JSON file and shared through the n2c2 community annotations repository [19]. Figure 2 illustrates a multianswer entry in the RxWhyQA data set.

Table 1. Categories, examples, and conversion strategies for making the drug-reason relations into the extractive question-answering annotations.

Category in the n2c2 ^a corpus	Example	Conversion strategy	Entries, n
1 Drug, no Reason	<i>Mirtazapine</i> 15 mg PO QHS ^b (only the drug is mentioned but no reason is documented)	Make an unanswerable QA ^c entry	46,278
1 Drug, 1 Reason	The patient received <i>morphine</i> for pain as needed	Make a 1-to-1 QA entry	28,224 ^d
N Drugs, 1 Reason	Hypertension: severely elevated blood pressure. Started <i>amlodipine</i> , <i>metoprolol</i> , and <i>isosorbide</i> .	Break into N separate 1-to-1 relations and make each a 1-to-1 QA entry	N/A ^e
1 Drug, N Reasons	<i>Albuterol sulfate</i> 90 mcg... Puff Inhalation Q4H ^f for sob or wheeze.	List the N reasons under the answer block to form a 1-to-N QA entry	22,437 ^g
M Drug, N Reasons	Left frontoparietal stroke - maintained on ASA ^h and <i>plavix</i> Hx of CVA ⁱ : restarted ASA/ <i>Plavix</i> per the GI ^j team's recommendation.	List the N reasons under answer block to form an M-to-N QA entry	N/A

^an2c2: National NLP (natural language processing) Clinical Challenges.

^bPO QHS: one pill to be taken orally at bedtime.

^cQA: question-answering.

^d28,224 entries in total for the 1-drug-1-reason and N-drugs-1-reason categories together in the corpus.

^eN/A: not applicable.

^fQ4H: every 4 hours.

^g22,437 entries in total for the 1-drug-N-reasons and M-drug-N-reasons categories in together in the corpus.

^hASA: acetylsalicylic acid (aspirin).

ⁱHx of CVA: history of cerebrovascular accident.

^jGI: gastrointestinal.

Table 2. Unique answers among answerable questions.

Frequency	Unique answers, n (%)
1	28,224 (75)
2	6804 (18)
3	1530 (4)
≥4	954 (3)

Figure 2. A multianswer entry in the generated RxWhyQA data set. The “id” field is the unique ID for the question-answering entry in the data set. The “_mname” field indicates the medication name; that is, the anchor concept in the question. The “answer_start” is the character offset where the answer term occurs in the clinical document, which is hosted in the “context” field (not shown here). When “is_impossible” is false, the question-answering entry is answerable.

```
{
  "question_template": "Why was the patient prescribed |medication|?",
  "question": "Why was the patient prescribed Metoprolol?",
  "id": "141586.xml_M100_3",
  "_mname": "Metoprolol",
  "answers": [
    {
      "text": "Atrial fibrillation",
      "answer_start": 8695
    },
    {
      "text": "Hypertension",
      "answer_start": 9115
    }
  ],
  "is_impossible": false
}
```


Content Analysis of the RxWhyQA Data Set

The 5 most frequently asked drug terms (with noting the number of QA entries) in the answerable questions (frequencies) were the following: coumadin (1278), vancomycin (1170), lasix (963), acetaminophen (801), and antibiotics (783). Without any overlap, the 5 most frequent drug terms in the unanswerable questions were the following: docusate sodium (648), metoprolol tartrate (504), aspirin (468), pantoprazole (450), and penicillins (414). Among the answerable QA entries, the 5 most frequently seen pairs were the following: acetaminophen-pain (504), senna-constipation (369), oxycodone-pain (261), coumadin-afib (252), and acetaminophen-fever (234). As a potential surrogate measure of task difficulty, Table 3 shows the distribution for

the number of sentences between the question anchor and answer term in each answerable QA entry. The majority (n=32,409, 72%) of the drug and reason terms occur within the same sentence, and the portion increases to 90% (72%+18%) when adding those with the drug and reason occurring in an adjacent sentence (ie, distance=1). In the extreme case, the drug and reason terms are 16 sentences apart from each other. Table 4 summarizes the commonly observed contexts from manually reviewing 100 random samples of the answerable QA entries. There were 7, 10, and 3 off-label uses, respectively, in each of the random 100 drug-reason pairs reviewed by the domain expert, making the estimate of off-label uses average at 6.7% in the RxWhyQA data set. The detailed off-label review results are available in Multimedia Appendix 1.

Table 3. Distribution for the distance between question and answer terms (0=the question and answer terms occur in the same sentence).

Distance (be sentence) between the question and answer items	QA ^a entries, n
0	32,409
1	8154
2	2646
3	1188
4	405
5	153
6	81
7	72
8	27
9	0
10	0
11	0
12	9
13	9
14	0
15	0
16	9

^aQA: question-answering.

Table 4. Common patterns (observed >10 times) between the question and the answer terms in 100 random question-answering entries. Each reason or drug represents where a question or answer anchor term occurs in the pattern. The shorthands are used as follows: ellipsis stands for 0 to multiple words, parentheses denote scoping, square brackets with pipes indicate a boolean OR set, and a question mark denotes a binary quantifier for presence or absence.

Pattern	Frequency
Reason ... (being)? [received started restarted required maintained continued?] (on)? Drug	25
Drug ... [prn PRN (as needed for)?] Reason	18
Drug ... (was)? [attempted given dosing taking] for (any)? [possible likely presumed]? Reason	14
Reason ... (was)? [managed treated improved recommended downtrended resolved reversed needed] with Drug	13

F1-Score of the Baseline EQA Solution

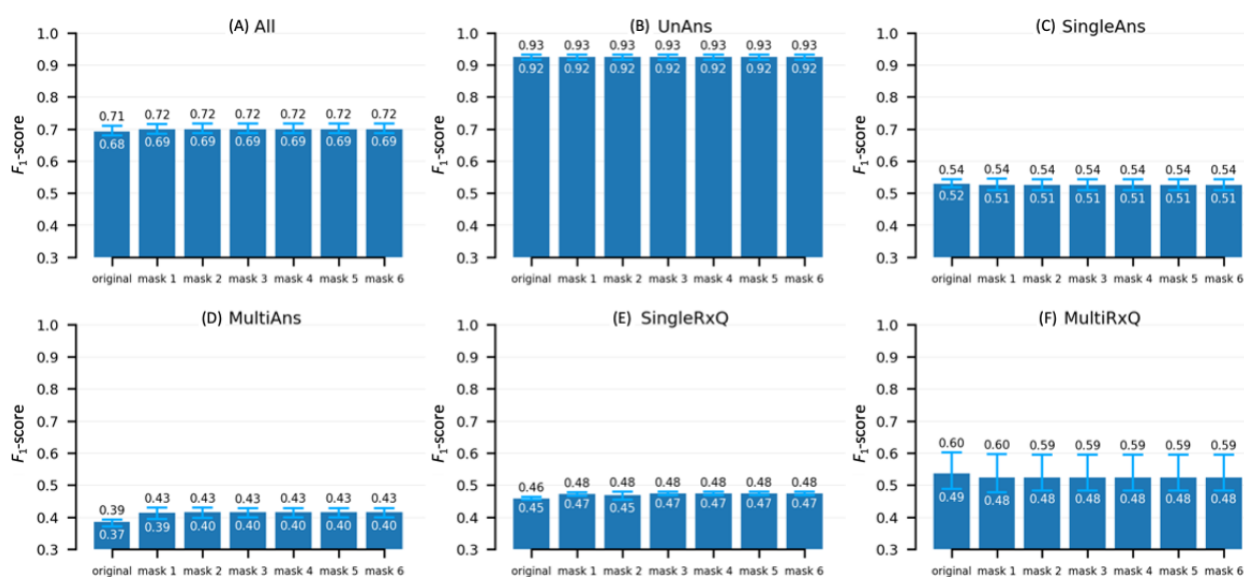
The performance in determining the F_1 -score across 3 experimental runs is summarized in Figure 3, where the subfigures represent different slices. Specifically, the underlying

set relations are the following: the full set (Figure 3A) minus the unanswerable questions (Figure 3B) yields the answerable questions, which can be represented by either single-answer questions (Figure 3C) plus multianswer questions (Figure 3D) if sliced per the number of answers or by questions asking about

a single drug (Figure 3E) plus questions asking about multiple drugs (Figure 3F) if sliced per the number of drugs asked in the question. Each bar represents the average F_1 -score across the runs and with the range marked for each incremental masking step. As seen in Figure 3A, the overall F_1 -score increased immediately after applying the first round of answer masking (from “original” to “mask 1”, $P < .05$), which then stayed constant throughout the remaining mask iterations. The increase in the F_1 -score in Figure 3A corresponds to the exact pattern in Figure 3D, suggesting that the performance gain was mainly from the multianswer questions; that is, the target originally intended by the masking. Multianswer questions appear to be more

challenging than single-answer questions on comparing Figures 3C and 3D. According to Figures 3E and 3F, asking about multiple drugs at once made it easier for the model to find the right answer, albeit with wide performance variation. The BERT model was good at refraining from answering unanswerable questions, as indicated by the high F_1 -scores in Figure 3B. The detailed results of the 3 experimental runs are available in Multimedia Appendix 2. There were 189 QA entries associated with the off-label uses identified by manually reviewing 300 random drug-reason pairs from the 3 test runs, all of which happened to be single-answer cases. We computed for this small set a single aggregate F_1 -score, which was 0.49 and appeared consistently lower than the range shown in Figure 3C.

Figure 3. F_1 -scores of the fine-tuned Bidirectional Encoder Representations from Transformers extractive question-answering model across the incremental masking rounds. Each bar represents the average F_1 -score based on 3 experimental runs, with the minimum and maximum range marked (light blue). (A) The full set, (B) unanswerable questions, (C) questions with exactly 1 answer, (D) questions with multiple answers, (E) questions asking about a single drug, and (F) questions asking about multiple drugs.



Discussion

Significance and Contributions

Although why-QA only covers a subdomain of clinical QA, it represents a unique category that deals with the cause, motivation, circumstance, and justification. It was estimated that 20% of the top 10 question types asked by family physicians [20] could be rephrased into a why-question. Clinical why-QA is important because (1) the ultimate task resembles expert-level explanatory synthesis of knowledge and evidence and (2) it aligns with identifying reasons for the decisions documented in clinical text. Therefore, the contents and challenges offered by the RxWhyQA data set itself have independent, practical value for developing clinical QA applications. Although drug-reason QA appears to be a niche topic, a working solution developed on the data set can broadly benefit research around adherence to clinical guidelines, care quality assessment, and health disparity from prescribing variations.

The generated RxWhyQA data set can serve as the training and testing of AI systems that target excerpting pertinent information in a clinical document to answer patient-specific questions. In

addition to the unanswerable questions that require a system to refrain from extracting FP answers, the RxWhyQA data set features 9288 questions that require the system to identify multiple answers, which is a realistic challenge in clinical QA. The data set also contains 611 questions that ask about the reason for prescribing multiple drugs at once. The multianswer and multifocus questions represent a key improvement beyond existing clinical EQA data sets, of which the rigid constructs would preclude AI solutions from learning to deal with more realistic use scenarios. Additionally, our experiments on these special constructs validated the challenging nature of multianswer questions and revealed that multifocus questions may turn out to be easier due to the availability of richer information for use by the model. Our drug-reason-focused data set may offer a coherent theme that enables better controlled experiments to compare how the different QA constructs (eg, single- vs multianswer questions) affect AI system performance.

Properties Found About the RxWhyQA Data Set

The frequent drugs and drug-reason pairs likely imply the clinical practice in the original n2c2 cohort. The finding that the top 5 drugs in the unanswerable questions (ie, no answer provided in the gold-standard annotation) were different from

those in the answerable questions suggests that the prescription of certain drugs might be self-evident without needing a documented reason. Our question-answer-mentioning distance analysis showed that 90% of the drug-reason pairs were within the same or an adjacent sentence in the RxWhyQA data set, indicating modest demand for long-distance inference by AI solutions. We were able to identify frequent contextual patterns such as “PRN” (ie, pro re nata) or “as needed for” (Table 4) that AI models may learn to facilitate locating the answers. It is estimated that the data set contains 6.7% of off-label drug uses as the target answers, which would be useful for training systems to identify such cases and facilitate research on understanding the medical practice variation or innovation.

Behavior of the Baseline EQA Solution

The notable increase in the F_1 -score (Figure 3D) after applying 1 round of masking suggests that the masking effectively forced the BERT model to look elsewhere, which resulted in an increase in the F_1 -score by retrieving the majority of the additional answers (see Table 2). Interestingly, we noticed in many cases that the model clung on to the masked span (ie, capturing the “_____” as an answer) where some of such strong contextual patterns were present. This phenomenon supports that transformer-based EQA models do leverage contextual information than merely memorizing the surface question-answer pairs. Moreover, our post hoc inspection noted that correct (synonymous) answers were found by the model that were not in the gold-standard annotation (eg, “allergic reaction” versus “anaphylaxis” to a question about “epipen”), suggesting that the performance could be underestimated. As a caveat, we were aware that our baseline solution was essentially a convenient hack that made a model trained for single-answer EQA to find multiple answers through a stepwise

probing procedure. As more advanced approaches constantly emerge [21,22], we welcome the research community to evaluate them by using the RxWhyQA data set. For example, the lower F_1 -score on those off-label uses indicates that they might represent challenging cases and demand more robust AI solutions.

Limitations

We admit several limitations in this study: (1) the source n2c2 corpus represented a specific cohort that may not generalize to every clinical data set, (2) we did not exhaustively diversify the paraphrastic questions but left it for future exploration on other promising approaches [23], (3) we did not intend to extensively compare state-of-the-art solutions for multianswer QA but rather intended to offer a convenience baseline along with releasing the RxWhyQA corpus, (4) the drug-reason relations represent a narrow topic for EQA development and evaluation. However, we believe that the definite theme would preferably make it a less confounded test set for assessing the effect of multianswer and multifocus questions on AI systems.

Conclusions

We derived and shared the RxWhyQA, an EQA data set for training and testing systems to answer patient-specific questions based on clinical documents. The RxWhyQA data set includes 9288 multianswer questions and 611 multifocus questions, each representing a critical scenario not well covered by existing data sets. Upon evaluating a baseline solution, the multianswer questions appeared to be more challenging than single-answer questions. Although the RxWhyQA focuses on why-questions derived from drug-reason relations, it offers a rich data set involving realistic constructs and exemplifies an innovation in recasting NLP annotations of different tasks for EQA research.

Acknowledgments

We thank the n2c2 organizers for making the annotations available to the research community. The study was partly supported by the Mayo Clinic Kern Center for the Science of Health Care Delivery. The research was supported by the National Center for Advancing Translational Sciences (U01TR002062).

Authors' Contributions

JWF conceived the study. HL offered scientific consultation. SM implemented the data conversion and analysis. HH assisted in the data conversion and graphic presentation. HJ reviewed and determined the off-label drug uses. SM and JWF drafted the manuscript. All authors contributed to the interpretation of the results and critical revision of the manuscript, and approved the final submission.

Conflicts of Interest

None declared.

Multimedia Appendix 1

Manual annotation of off-label uses in 300 randomly sampled drug-reason QA pairs from the test sets.

[[XLSX File \(Microsoft Excel File\), 40 KB - ai_v2i1e41818_app1.xlsx](#)]

Multimedia Appendix 2

Detailed F_1 -scores of the BERT model across three test runs, on the different subsets, with applying the incremental answer-masking.

[[XLSX File \(Microsoft Excel File\), 13 KB - ai_v2i1e41818_app2.xlsx](#)]

References

1. Cimino J. Putting the "why" in "EHR": capturing and coding clinical cognition. *J Am Med Inform Assoc* 2019 Nov 01;26(11):1379-1384 [FREE Full text] [doi: [10.1093/jamia/ocz125](https://doi.org/10.1093/jamia/ocz125)] [Medline: [31407781](https://pubmed.ncbi.nlm.nih.gov/31407781/)]
2. Goodwin T, Harabagiu S. Medical question answering for clinical decision support. 2016 Presented at: CIKM'16: ACM Conference on Information and Knowledge Management; October 24-28, 2016; Indianapolis, IN. [doi: [10.1145/2983323.2983819](https://doi.org/10.1145/2983323.2983819)]
3. Jin Q, Yuan Z, Xiong G, Yu Q, Ying H, Tan C, et al. Biomedical question answering: a survey of approaches and challenges. *ACM Comput Surv* 2022 Jan 18;55(2):1-36. [doi: [10.1145/3490238](https://doi.org/10.1145/3490238)]
4. Yue X, Gutierrez BJ, Sun H. Clinical reading comprehension: a thorough analysis of the emrQA dataset. *arXiv Preprint* posted online May 1, 2020. [doi: [10.18653/v1/2020.acl-main.410](https://doi.org/10.18653/v1/2020.acl-main.410)]
5. Yoon W, Jackson R, Lagerberg A, Kang J. Sequence tagging for biomedical extractive question answering. *Bioinformatics* 2022 Aug 02;38(15):3794-3801 [FREE Full text] [doi: [10.1093/bioinformatics/btac397](https://doi.org/10.1093/bioinformatics/btac397)] [Medline: [35713500](https://pubmed.ncbi.nlm.nih.gov/35713500/)]
6. McCann B, Keskar NS, Xiong C, Socher R. The natural language decathlon: multitask learning as question answering. *arXiv Preprint* posted online June 20, 2018.
7. Wang L, Zheng K, Qian L, Li S. A survey of extractive question answering. 2022 Presented at: 2022 International Conference on High Performance Big Data and Intelligent Systems (HDIS); December 10-11, 2022; Tianjin, China. [doi: [10.1109/hdis56859.2022.9991478](https://doi.org/10.1109/hdis56859.2022.9991478)]
8. Rajpurkar P, Zhang J, Lopyrev K, Liang P. SQuAD: 100,000+ questions for machine comprehension of text. In: *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*. 2016 Presented at: 2016 Conference on Empirical Methods in Natural Language Processing; 2016; Austin, TX p. 2383-2392. [doi: [10.18653/v1/d16-1264](https://doi.org/10.18653/v1/d16-1264)]
9. Rajpurkar P, Jia R, Liang P. Know what you don't know: unanswerable questions for SQuAD. In: *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*. 2018 Presented at: 56th Annual Meeting of the Association for Computational Linguistics; 2018; Melbourne p. 784-789. [doi: [10.18653/v1/p18-2124](https://doi.org/10.18653/v1/p18-2124)]
10. Raghavan P, Patwardhan S, Liang J, Devarakonda M. Annotating electronic medical records for question answering. *arXiv Preprint* posted online May 17, 2018. [doi: [10.48550/arXiv.1805.06816](https://doi.org/10.48550/arXiv.1805.06816)]
11. Zhu M, Ahuja A, Juan DC, Wei W, Reddy CK. Question answering with long multiple-span answers. 2020 Presented at: The 2020 Conference on Empirical Methods in Natural Language Processing; November 16-20, 2020; Online. [doi: [10.18653/v1/2020.findings-emnlp.342](https://doi.org/10.18653/v1/2020.findings-emnlp.342)]
12. Ju Y, Wang W, Zhang Y, Zheng S, Liu K, Zhao J. CMQA: A Dataset of Conditional Question Answering with Multiple-Span Answers. 2022 Presented at: 29th International Conference on Computational Linguistics; 2022; Gyeongju, Republic of Korea.
13. Pampari A, Raghavan P, Liang J, Peng J. emrQA: A large corpus for question answering on electronic medical records. 2018 Presented at: The 2018 Conference on Empirical Methods in Natural Language Processing; October 31-November 4, 2018; Brussels, Belgium. [doi: [10.18653/v1/d18-1258](https://doi.org/10.18653/v1/d18-1258)]
14. Henry S, Buchan K, Filannino M, Stubbs A, Uzuner O. 2018 n2c2 shared task on adverse drug events and medication extraction in electronic health records. *J Am Med Inform Assoc* 2020 Jan 01;27(1):3-12 [FREE Full text] [doi: [10.1093/jamia/ocz166](https://doi.org/10.1093/jamia/ocz166)] [Medline: [31584655](https://pubmed.ncbi.nlm.nih.gov/31584655/)]
15. Moon SR, Fan J. How you ask matters: the effect of paraphrastic questions to BERT performance on a clinical SQuAD dataset. 2020 Presented at: The 3rd Clinical Natural Language Processing Workshop; November 19, 2020; Online. [doi: [10.18653/v1/2020.clinicalnlp-1.13](https://doi.org/10.18653/v1/2020.clinicalnlp-1.13)]
16. Zheng NS, Kerchberger VE, Borza VA, Eken HN, Smith JC, Wei W. An updated, computable MEDication-Indication resource for biomedical research. *Sci Rep* 2021 Sep 23;11(1):18953 [FREE Full text] [doi: [10.1038/s41598-021-98579-4](https://doi.org/10.1038/s41598-021-98579-4)] [Medline: [34556781](https://pubmed.ncbi.nlm.nih.gov/34556781/)]
17. Devlin J, Chang M, Lee K, Toutanova K. BERT: pre-training of deep bidirectional transformers for language understanding. *arXiv Preprint* posted online October 11, 2018. [doi: [10.48550/arXiv.1810.04805](https://doi.org/10.48550/arXiv.1810.04805)]
18. Alsentzer E, Murphy J, Boag W, Weng WH, Jindi D, Naumann T, et al. Publicly Available Clinical BERT Embeddings. 2019 Presented at: The 2nd Clinical Natural Language Processing Workshop; June 7, 2019; Minneapolis, MN.
19. n2c2 Data Upload: Community generated annotations. DBMI Data Portal. URL: <https://portal.dbmi.hms.harvard.edu/projects/n2c2-du/> [accessed 2023-05-29]
20. Ely JW, Osheroff JA, Ebell MH, Bergus GR, Levy BT, Chambliss ML, et al. Analysis of questions asked by family doctors regarding patient care. *BMJ* 1999 Aug 07;319(7206):358-361 [FREE Full text] [doi: [10.1136/bmj.319.7206.358](https://doi.org/10.1136/bmj.319.7206.358)] [Medline: [10435959](https://pubmed.ncbi.nlm.nih.gov/10435959/)]
21. Hu M, Peng Y, Huang Z, Li D. A multi-type multi-span network for reading comprehension that requires discrete reasoning. *arXiv Preprint* posted online August 15, 2019. [doi: [10.18653/v1/d19-1170](https://doi.org/10.18653/v1/d19-1170)]
22. Segal E, Efrat A, Shoham M, Globerson A, Berant J. A Simple and Effective Model for Answering Multi-span Questions. 2020 Presented at: The 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP); November 16-20, 2020; Online. [doi: [10.18653/v1/2020.emnlp-main.248](https://doi.org/10.18653/v1/2020.emnlp-main.248)]
23. Soni S, Roberts K. A paraphrase generation system for EHR question answering. 2019 Presented at: The 18th BioNLP Workshop and Shared Task; August 1, 2019; Florence, Italy. [doi: [10.18653/v1/w19-5003](https://doi.org/10.18653/v1/w19-5003)]

Abbreviations

AI: artificial intelligence
BERT: Bidirectional Encoder Representations from Transformers
dev: develop
EHR: electronic health record
EQA: extractive question-answering
FN: false negative
FP: false positive
n2c2: National NLP Clinical Challenges
NLP: natural language processing
QA: question-answering
SQuAD: Stanford Question Answering Dataset
TP: true positive

Edited by K El Emam, B Malin; submitted 09.08.22; peer-reviewed by Z Yin, I Danciu; comments to author 02.11.22; revised version received 31.01.23; accepted 22.05.23; published 20.06.23.

Please cite as:

Moon S, He H, Jia H, Liu H, Fan JW

Extractive Clinical Question-Answering With Multianswer and Multifocus Questions: Data Set Development and Evaluation Study
JMIR AI 2023;2:e41818

URL: <https://ai.jmir.org/2023/1/e41818>

doi: [10.2196/41818](https://doi.org/10.2196/41818)

PMID: [38875580](https://pubmed.ncbi.nlm.nih.gov/38875580/)

©Sungrim Moon, Huan He, Heling Jia, Hongfang Liu, Jungwei Wilfred Fan. Originally published in JMIR AI (<https://ai.jmir.org>), 20.06.2023. This is an open-access article distributed under the terms of the Creative Commons Attribution License (<https://creativecommons.org/licenses/by/4.0/>), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR AI, is properly cited. The complete bibliographic information, a link to the original publication on <https://www.ai.jmir.org/>, as well as this copyright and license information must be included.

Original Paper

Momentary Depressive Feeling Detection Using X (Formerly Twitter) Data: Contextual Language Approach

Ali Akbar Jamali¹, PhD; Corinne Berger¹, PhD; Raymond J Spiteri¹, PhD

Department of Computer Science, University of Saskatchewan, Saskatoon, SK, Canada

Corresponding Author:

Ali Akbar Jamali, PhD

Department of Computer Science

University of Saskatchewan

S415 Thorvaldson Building

110 Science Place

Saskatoon, SK, S7N5C9

Canada

Phone: 1 306 966 2925

Email: A.A.Jamali@usask.ca

Abstract

Background: Depression and momentary depressive feelings are major public health concerns imposing a substantial burden on both individuals and society. Early detection of momentary depressive feelings is highly beneficial in reducing this burden and improving the quality of life for affected individuals. To this end, the abundance of data exemplified by X (formerly Twitter) presents an invaluable resource for discerning insights into individuals' mental states and enabling timely detection of these transitory depressive feelings.

Objective: The objective of this study was to automate the detection of momentary depressive feelings in posts using contextual language approaches.

Methods: First, we identified terms expressing momentary depressive feelings and depression, scaled their relevance to depression, and constructed a lexicon. Then, we scraped posts using this lexicon and labeled them manually. Finally, we assessed the performance of the Bidirectional Encoder Representations From Transformers (BERT), A Lite BERT (ALBERT), Robustly Optimized BERT Approach (RoBERTa), Distilled BERT (DistilBERT), convolutional neural network (CNN), bidirectional long short-term memory (BiLSTM), and machine learning (ML) algorithms in detecting momentary depressive feelings in posts.

Results: This study demonstrates a notable distinction in performance between binary classification, aimed at identifying posts conveying depressive sentiments and multilabel classification, designed to categorize such posts across multiple emotional nuances. Specifically, binary classification emerges as the more adept approach in this context, outperforming multilabel classification. This outcome stems from several critical factors that underscore the nuanced nature of depressive expressions within social media. Our results show that when using binary classification, BERT and DistilBERT (pretrained transfer learning algorithms) may outperform traditional ML algorithms. Particularly, DistilBERT achieved the best performance in terms of area under the curve (96.71%), accuracy (97.4%), sensitivity (97.57%), specificity (97.22%), precision (97.30%), and F_1 -score (97.44%). DistilBERT obtained an area under the curve nearly 12% points higher than that of the best-performing traditional ML algorithm, convolutional neural network. This study showed that transfer learning algorithms are highly effective in extracting knowledge from posts, detecting momentary depressive feelings, and highlighting their superiority in contextual analysis.

Conclusions: Our findings suggest that contextual language approaches—particularly those rooted in transfer learning—are reliable approaches to automate the early detection of momentary depressive feelings and can be used to develop social media monitoring tools for identifying individuals who may be at risk of depression. The implications are far-reaching because these approaches stand poised to inform the creation of social media monitoring tools and are pivotal for identifying individuals susceptible to depression. By intervening proactively, these tools possess the potential to slow the progression of depressive feelings, effectively mitigating the societal load of depression and fostering improved mental health. In addition to highlighting the capabilities of automated sentiment analysis, this study illuminates its pivotal role in advancing global public health.

(JMIR AI 2023;2:e49531) doi:[10.2196/49531](https://doi.org/10.2196/49531)

KEYWORDS

depression; momentary depressive feelings; X (Twitter); natural language processing; lexicon; machine learning; transfer learning

Introduction

Mental health is an essential aspect of people's overall well-being and daily functioning. According to the World Health Organization [1], approximately 25% of the global population experiences a mental health condition at some point in their life, making mental disorders a significant public health concern. Mental health conditions can have substantial socioeconomic impacts on individuals and society, including reduced quality of life and workforce productivity and increased health care costs. Accordingly, it is vitally important to prioritize and address mental health conditions by adopting effective strategies to curb their prevalence [2].

Subthreshold depression, or momentary depressive feelings, refers to depression symptoms that are not severe enough to be considered as a major depressive disorder [3]. Despite not meeting the criteria for a depression diagnosis, momentary depressive feelings can still significantly impact an individual's daily life and well-being. Frequent or prolonged experiences of momentary depressive feelings can be a sign of the development of depression [4-6] and can lead to decreased energy, loss of interest in activities, and persistent low mood [7-9]. Mitchell et al [10] found that people with subthreshold depression reported higher levels of disability and decreased quality of life compared to those not reporting any symptoms.

Early detection of momentary depressive feelings is important for an individual's mental health and well-being because it allows identifying individuals who may be at a higher risk of developing depression [11,12] and lets them address these feelings before they escalate into a more severe form of depression [6,13]. Detecting momentary depressive feelings can also provide important information for researchers and mental health professionals, leading to a better understanding of the nature of depression, and can be used to provide preventative interventions and support to help individuals maintain their mental health. However, it is worth noting that depression is a complex condition with multiple causes, and the detection of momentary depressive feelings should be considered in conjunction with a comprehensive evaluation of an individual's overall mental health [14].

Momentary depressive feelings can be detected through different standard methods including self-report measures, behavioral observations, and physiological measures [15]. Self-report measures ask individuals to reflect on their current mood and symptoms, whereas behavioral observations involve observing and recording an individual's behavior and facial expressions. Physiological measures, such as measuring cortisol levels, heart rate variability, and skin conductance, can also provide insight into an individual's emotional state [16].

Momentary depressive feelings are often accompanied by distinctive linguistic patterns, allowing us to understand an individual's emotional state and cognitive processes. One prevalent symptom is the expression of negative sentiment and emotion [17]. People with momentary depressive feelings or

who are in depression frequently use language dominated by a pessimistic lexicon, conveying feelings of hopelessness, sadness, and despair. This linguistic tendency reflects their internal emotional turmoil and offers an insight into the depth of their distress. Another linguistic hallmark is the increase in self-referential language. The excessive use of first-person pronouns such as "I" or "me" suggests a potential focus on one's own experiences and an emphasis on the self [18,19]. The study of these linguistic symptoms within contextual contents offers promising avenues for the detection of momentary depressive feelings using advanced methods. Contextual approaches have demonstrated outstanding ability in discerning linguistic markers of depression. These methods consider not only individual words but also the surrounding context, enabling a more accurate interpretation of the intended meaning.

One prominent opportunity in this domain involves sentiment analysis. Sentiment analysis attempts to gauge the emotional tone of the text. Contextual approaches for sentiment analysis such as Valence Aware Dictionary for Sentiment Reasoning (VADER) rely on predefined sentiment scores assigned to words. These models can swiftly identify texts with predominantly negative sentiments [20]. However, their generic lexicons might not capture the nuanced depressive expressions. Machine learning (ML) techniques such as support vector machines and decision trees have also been applied to classify depressive textual content. These models learn to differentiate between depressive and nondepressive contents by extracting features from the text, including n-grams and linguistic patterns [21]. Yet, these methods can struggle with complex contextual cues inherent in sentiment discourse. Subsequently, more sophisticated models have emerged using contextual methods such as Word2Vec and FastText to discover semantic relationships within text. These models offer the advantage of representing words in context, which is vital for detecting subtler depressive symptoms [22]. Similarly, deep learning architectures such as convolutional neural network (CNN) and long short-term memory network have been used to capture sequential dependencies in text, enhancing the understanding of the temporal progression of depressive feelings [23,24].

In recent years, the advent of pretrained language models such as Bidirectional Encoder Representations From Transformers (BERT) [25] has revolutionized the application of natural language processing (NLP) [26], including depressive feeling detection. NLP offers promising alternative methods for the detection of momentary depressive feelings using social media, such as Facebook, Instagram, and X (formerly Twitter), where individuals can broadcast their thoughts and feelings in real time [27]. NLP can be used to identify patterns in language and sentiment to recognize specific language and behavioral markers that may be indicative of depression. Sentiment analysis can be used to determine the underlying emotional tone and identify individuals at risk of depression using large-scale data sets. Numerous studies have applied sentiment analysis and text classification in the area of mental health [28-30] to detect effectively depressive feelings and depression [31-36].

X has emerged as a popular social media platform for mental health research due to its vast and diverse text-based data. Its real-time nature makes it particularly well-suited for studying a variety of mental disorders such as dementia [37], depression [38], Alzheimer disease [39], and schizophrenia [40]. The analysis of X data allows the detection of momentary mood changes and depressive feelings, offering a unique opportunity for identifying individuals who may be at risk of developing depression. X data also provide a chance to investigate the expression of mental disorders and develop novel ML-based methods for detecting other mental disorders.

The primary objective of this study is to develop contextual language approaches and assess their effectiveness in detecting momentary depressive feelings in posts. With the aid of large-scale data and advanced NLP techniques, this study aims to analyze linguistic features and patterns in posts to detect momentary depressive feelings. This study has the potential to provide valuable insights into the relationship between language

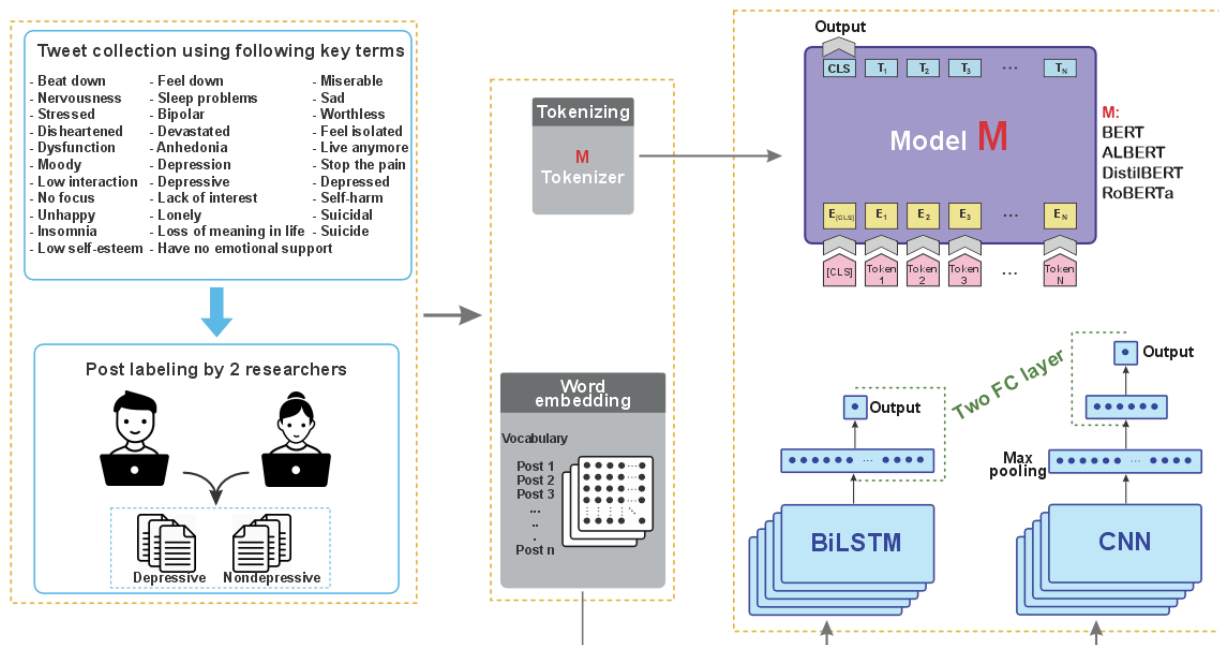
and mental health. Additionally, it may contribute to the development of tools for the early detection and prevention of depression.

Methods

Study Design

The overview and workflow of this study are presented in Figure 1. First, we identified and collected words expressing depression. We then scraped and manually labeled posts. Next, the posts in our data set were preprocessed and cleaned. Finally, state-of-the-art NLP algorithms were used for the detection of momentary depressive feelings in posts. This study was performed using Python (version 3.9; Python Software Foundation) and R (R Foundation for Statistical Computing) programming languages with different packages. The data collection, preparation, and ML algorithms used in this study are discussed in the following sections.

Figure 1. The workflow of data collection and architecture of the proposed detection approaches. ALBERT: A Lite BERT; BERT: Bidirectional Encoder Representations From Transformers; BiLSTM: Bidirectional Long Short-term Memory; CLS: classification tasks; CNN: Convolutional Neural Network; DistilBERT: Distilled BERT; FC: fully connected; RoBERTa: Robustly Optimized BERT Approach.



Data Collection and Preparation

Overview

Contextual language approaches have been found to be effective in processing large text data sets, making them appropriate for text-based tasks such as sentiment analysis and for determining whether individual posts may express momentary depressive feelings. Accordingly, in order to collect appropriate data, we first need to construct a suitable lexicon and then use it to scrape for appropriate posts. The data are then manually labeled in preparation for training the ML algorithms.

Lexicon Construction

In the process of lexicon construction, we carefully examined and reviewed major studies that delve into the correlation

between social media and depression, aiming to compile essential terms for our depression lexicon [41,42]. We specifically focused on key terms relevant to depressive feelings and collected a variety of these terms as previously highlighted in relevant research. Following the elimination of any redundant terms, we arrived at a total of 41 distinct key terms within our lexicon. Recognizing that the quality of the lexicon is pivotal for accurate contextual language analysis, 2 researchers (AAJ and CB) assessed the relevance of each term in our lexicon to depressive feelings using a 5-point scale, where a rating of 5 indicates significant relevance. In this study, we elected to only incorporate key terms with a score of 3 or higher. By applying this criterion, the lexicon was refined to contain 32 terms, as illustrated in Textbox 1.

Textbox 1. The lexicon for momentary depressive feeling detection.

Included depressive feeling lexicons (n=32)

- Beat down, nervousness, stressed, disheartened, dysfunction, moody, low interaction, no focus, unhappy, insomnia, low self-esteem, feel down, have no emotional support, sleep problems, bipolar, devastated, anhedonia, depression, depressive, lack of interest, lonely, miserable, sad, worthless, feel isolated, live anymore, stop the pain, depressed, loss of meaning in life, self-harm, suicidal, and suicide

Excluded depressive feeling lexicons (n=9)

- Instability, imbalance, broken, disillusioned, emotional, uneasiness, disturbed, anxiety, and fatigue

Post Scraping

The *Twint* package was used to collect posts using a constructed lexicon and key terms. *Twint* is a Python package that allows scraping posts, followings, followers, likes, etc, from X without using X's application programming interface. It provides several customization options for searching specific posts based on keywords, location, language, date, and more. Unlike the X application programming interface, *Twint* does not have any rate limits or access restrictions, making it possible to scrape large amounts of data. posts posted between January 1, 2022, and December 30, 2022, were extracted.

Post Labeling

Two researchers (AAJ and CB) labeled posts as expressing or not expressing momentary depressive feelings. To determine intercoder reliability, the researchers read and scored the same 100 posts on a 5-point scale, with 5 indicating significant relevance to momentary depressive feelings, and achieved intercoder reliability of 86% based on the percent agreement method [43]. Each researcher was then given 2000 posts to label. For binary classification, out of 4000 labeled posts, 1840 posts that received a score of 3 or higher were used as positive samples. To construct a balanced data set, 1840 negative samples (posts not expressing momentary depressive feelings) were randomly selected, and a final data set was constructed consisting of 3680 samples with an equal number of positive and negative samples. For multilabel classification, all labeled posts in 5 categories on a scale of 1 to 5 were included.

Preprocessing

Given that posts frequently contain misspelled words, irrelevant characters, emoticons, and unconventional syntax—considered noise in NLP—we implemented various preprocessing steps on our data set before training the algorithms. Key preprocessing steps included:

1. Filtration: Removing punctuations, emoticons, duplicates, replies, URLs, and HTML links
2. Tokenization: Breaking a phrase or sentence into individual words called tokens
3. Lowercasing: Converting all uppercase letters to lowercase and ensuring consistent word vectors across multiple instances of the same word
4. Lemmatization: Removing inflectional parts in a word or converting the word into its base form
5. Stemming: Removing prefixes or suffixes from words to obtain their root form
6. Removing stop words: Articles, prepositions, and pronouns (eg, the, in, a, an, and with) known as the “stop words” are

uninformative, and removing them helps the model to focus on the important words

ML Algorithms

Transfer Learning Algorithms

BERT Algorithm

BERT [25] is a state-of-the-art deep learning algorithm used extensively for NLP tasks using a transformer architecture to learn contextual representations of words in a sentence. BERT scrutinizes the words and their relationships in the post to capture nuanced meanings and emotions to analyze the textual content of the posts (eg, posts expressing momentary depressive feelings).

BERT Family Algorithms

In this study, we also used several subtypes of the BERT algorithm, that is, A Lite BERT (ALBERT) [44], Robustly Optimized BERT Approach (RoBERTa) [45], and Distilled BERT (DistilBERT) [46], to investigate their performance in detecting momentary depressive feelings.

Traditional ML Algorithms

CNN Algorithm

The CNN [47] is a deep learning algorithm that has been widely used in various computer vision and, more recently, in text analysis tasks. In post classification, CNNs can be used to detect posts with particular content (eg, momentary depressive feelings) by learning a representation of the post's text that is fed into an algorithm to make a detection [48,49].

Bidirectional Long Short-Term Memory

Bidirectional long short-term memory (BiLSTM) is a type of recurrent neural network that is particularly well suited for processing sequential data such as text [50]. Unlike traditional neural networks that process input sequences in only one direction, BiLSTMs process the input sequence in both forward and backward directions, enabling the network to capture contextual information from both past and future time steps. This capability results in improved performance on text mining tasks. In the context of post analysis, a BiLSTM can be trained on a large data set of posts to learn the contextual relationships between the words in a post and detect whether a post expresses momentary depressive feelings. The BiLSTM considers the order of words in a post and the relationships between them, enabling it to capture more complex patterns and relationships in the data compared to traditional feedforward neural network.

Evaluation

To investigate and evaluate the performance of competing algorithms, a 10-fold cross-validation (CV) approach is carried out. This involved randomly partitioning the input data into 2 sets: a CV set (2944/3680, 80% labeled posts) and a test set (736/3680, 20% labeled posts). The CV set was further divided into 10 subsets, allowing us to construct and train 10 distinct models. These models were subsequently evaluated using unseen data. The use of unseen data in CV is crucial for assessing the generalization capability of the models. This evaluation with unseen data helps mitigate the risk of overfitting and provides a more reliable estimation of the algorithm's performance in real-world scenarios [51]. The algorithm's overall performance was computed by averaging the results obtained from the 10 runs. To assess the performance of competing algorithms to

accurately identify posts expressing momentary depressive feelings, 6 evaluation metrics were calculated: area under the curve (AUC), accuracy, sensitivity, specificity, precision, and F_1 -score [52].

Hyperparameter Sensitivity

To optimize algorithm performance, the tuning of hyperparameters emerges as a necessity. Yet, it is important to recognize that a universal approach for hyperparameter selection is not known to exist. In light of this, this study took a multifaceted approach by delving into distinct sets of hyperparameter values for each algorithm. This approach enabled us to carefully identify the configuration that attains high performance. A summary of contextual approaches and their respective hyperparameters used in this study can be found in [Table 1](#).

Table 1. Hyperparameters for different contextual approaches.

Algorithm	Hyperparameters
BERT ^a , ALBERT ^b , RoBERTa ^c , and DistilBERT ^d	<ul style="list-style-type: none"> Learning rate (Adam): (5×10^{-5}, 4×10^{-5}, 3×10^{-5}, 2×10^{-5}, and 1×10^{-5}) Batch size: (8, 16, 32, and 64) Training epochs: (2, 3, 4, 5, 6, and 7)
CNN ^e	<ul style="list-style-type: none"> Learning rate: (1×10^{-4}, 1×10^{-3}, and 1×10^{-2}) Kernel size: (2, 3, 4, 5, and 6) Batch size: (16, 32, 64, and 128) Training epochs: (2, 3, 4, 5, 6, and 7)
BiLSTM ^f	<ul style="list-style-type: none"> Batch size: (16, 32, 64, and 128) Training epochs: (2, 3, 4, 5, 6, and 7)

^aBERT: Bidirectional Encoder Representations From Transformers.

^bALBERT: A Lite BERT.

^cRoBERTa: Robustly Optimized BERT Approach.

^dDistilBERT: Distilled BERT.

^eCNN: convolutional neural network.

^fBiLSTM: bidirectional long short-term memory.

Ethical Considerations

In contrast to conventional research involving human participants, ethical guidelines pertaining to social media research propose that publicly accessible data (eg, posts publicly posted on X) can be used for research purposes without necessitating supplementary consent or ethics endorsement [53,54]. In this study, we did not interact and intervene with the users whose public posts were collected and analyzed user-generated posts. It is worth noting, however, that any potentially associated identifying personal information (eg, user

IDs and URLs) has been carefully eliminated to uphold anonymity and safeguard the privacy of X users.

Results

Hyperparameter Sensitivity Analysis

In this study, various contextual language approaches were used, each with a range of hyperparameters tuned to achieve optimal performance. Multiple iterations of each algorithm were carried out using different hyperparameter configurations. The hyperparameter combinations that yielded the highest performance for each model are summarized in [Table 2](#).

Table 2. Optimal hyperparameter configurations.

Algorithm	Hyperparameters			
	Learning rate (Adam)	Batch size	Training epochs	Kernel size
BERT ^a	3×10^{-5}	16	3	N/A ^b
ALBERT ^c	2×10^{-5}	16	4	N/A
RoBERTa ^d	1×10^{-5}	32	3	N/A
DistilBERT ^e	2×10^{-5}	16	3	N/A
CNN ^f	1×10^{-3}	64	4	4
BiLSTM ^g	N/A	32	3	N/A

^aBERT: Bidirectional Encoder Representations From Transformers.

^bN/A: not applicable.

^cALBERT: A Lite BERT.

^dRoBERTa: Robustly Optimized BERT Approach.

^eDistilBERT: Distilled BERT.

^fCNN: convolutional neural network.

^gBiLSTM: bidirectional long short-term memory.

Performance Assessment

In this study, we undertook both binary and multilabel classifications. Nevertheless, it is noteworthy that the outcomes of the multilabel classification (see [Multimedia Appendix 1](#) for details) were not as encouraging as those achieved in the binary classification task. Our analysis revealed an interesting finding in the context of multilabel classification, namely, that posts expressing depressive feelings, regardless of their intensity or scale, pose a challenge for classification models. The complexity of these posts makes it challenging to achieve precise categorization since they encompass a range of emotional states that may not align neatly with predefined categories. This insight highlights the complexity of classifying nuanced sentiment, particularly in the context of depressive expressions.

The performance of different algorithms in detecting momentary depressive feelings with binary classification is presented in [Table 3](#). Our results indicated that BERT and DistilBERT outperformed in momentary depressive feelings detection and achieved the highest values in almost all performance metrics with AUC values of 95.80% and 96.71%, respectively. Additionally, both algorithms demonstrated high accuracy (96.03% and 97.40%), sensitivity (96.22% and 97.57%), specificity (95.83% and 97.22%), precision (95.96% and 97.30%), and F_1 -score (96.09% and 97.44%). The performance of traditional ML algorithms was relatively poor with the highest scores achieved by CNN (AUC: 84.81% and accuracy: 84.79%) and BiLSTM (AUC: 79.91% and accuracy: 79.86%). These findings indicated that the transfer learning algorithms performed significantly superior by a substantial margin. For instance, DistilBERT achieved an AUC value nearly 12% points

higher than the highest AUC achieved by CNN (84.81%). These findings confirm the feasibility of this algorithm in detecting momentary depressive feelings highlighting the effectiveness of transfer learning algorithms in NLP tasks.

It is important to note that the transfer learning algorithms, especially DistilBERT and BERT, achieved high values in other performance metrics such as sensitivity, specificity, precision, and F_1 -score, in addition to overall accuracy. High sensitivity and specificity demonstrate that these algorithms were able to accurately identify posts with momentary depressive feelings while avoiding false positive and false negative predictions. The significant performance variation observed between BERT and its more lightweight counterpart, ALBERT, was an important finding of this study. ALBERT incorporates parameter-reduction techniques, which may impact its ability to capture intricate nuances within the data as effectively as BERT. Furthermore, we have closely examined potential disparities in pretraining strategies and fine-tuning procedures, seeking to identify any factors that might contribute to the observed divergence in performance. By elaborating on these architectural and procedural distinctions, we aim to provide a comprehensive understanding of the reasons underlying BERT's superior performance over ALBERT. This analysis not only informs this study but also contributes to the broader discourse on the comparative strengths and limitations of these prominent language models.

Overall, these results support the use of transfer learning algorithms for momentary depressive feelings detection in posts. [Figure 2](#) depicts the class-wise results of competing algorithms using confusion matrices.

Table 3. The performance of different algorithms using post binary classification.

Algorithm	Performance metrics					
	AUC ^a (%)	Accuracy (%)	Sensitivity (%)	Specificity (%)	Precision (%)	F ₁ -score (%)
BERT ^b	95.80	96.03	96.22	95.83	95.96	96.09
ALBERT ^c	81.36	86.71	87.84	85.56	86.21	87.01
RoBERTa ^d	84.15	84.25	93.22	75.07	79.26	85.68
DistilBERT ^e	<i>96.71</i> ^f	<i>97.40</i>	<i>97.57</i>	<i>97.22</i>	<i>97.30</i>	<i>97.44</i>
CNN ^g	84.81	84.79	77.78	91.97	90.82	83.80
BiLSTM ^h	79.91	79.86	75.34	84.49	83.23	79.09

^aAUC: area under the curve.

^bBERT: Bidirectional Encoder Representations From Transformers.

^cALBERT: A Lite BERT.

^dRoBERTa: Robustly Optimized BERT Approach.

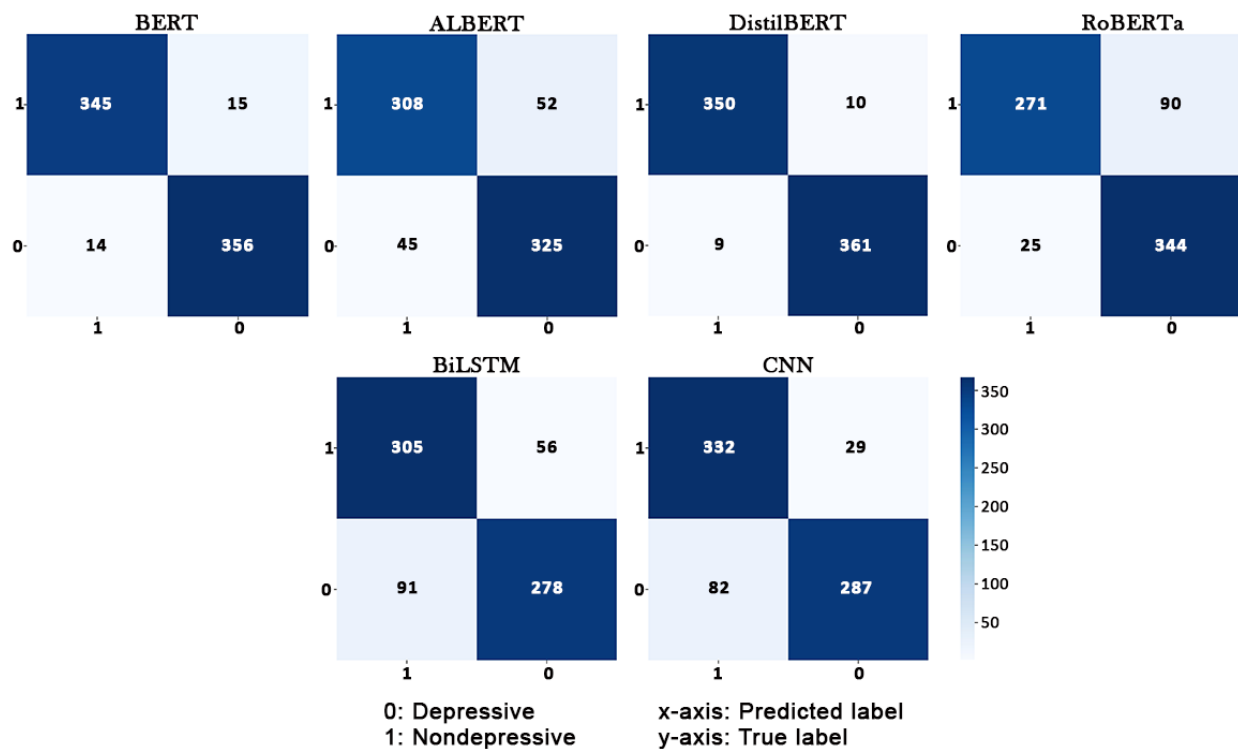
^eDistilBERT: Distilled BERT.

^fThe best values for the performance metrics are in italics.

^gCNN: convolutional neural network.

^hBiLSTM: bidirectional long short-term memory.

Figure 2. Confusion matrices produced by the competing algorithms for the test set. ALBERT: A Lite BERT; BERT: Bidirectional Encoder Representations From Transformers; BiLSTM: bidirectional long short-term memory; CNN: convolutional neural network; DistilBERT: Distilled BERT; RoBERTa: Robustly Optimized BERT Approach.



Discussion

Principal Findings

This study aimed to detect momentary depressive feelings in X data using contextual language approaches. Our results indicated that (pretrained) transfer learning algorithms such as DistilBERT can effectively detect momentary depressive feelings with an

accuracy of 97.4%. In this study, we used a comprehensive process of lexicon construction, data collection, and NLP-based ML algorithms to obtain accurate results. Our findings have practical implications in the mental health field by offering a potential framework for monitoring and detecting individuals' mental health states in real time, which could facilitate timely interventions and support. For instance, this research can contribute to the development of automated systems that analyze

social media posts, enabling mental health professionals to identify individuals expressing depressive feelings and subsequently provide support and resources.

The success of this study can be largely attributed to the effective use of a precise lexicon that contained a list of key terms relevant to depressive feelings based on prior research. These key terms were then manually evaluated to ensure their accuracy and relevance. The high intercoder reliability achieved by the researchers is a testament to the quality of the post labeling. This approach fostered the quality of the lexicon, leading to the accurate detection of momentary depressive feelings. Furthermore, the manual labeling of posts also contributed to the accuracy of the study because automatic labeling typically introduces noise into data and degrades algorithm performance [55].

In this study, we used a range of algorithms, including BERT, ALBERT, RoBERTa, DistilBERT, CNN, and BiLSTM. Notably, DistilBERT outperformed the other algorithms in detecting momentary depressive feelings. The use of multiple algorithms allowed for a comprehensive evaluation of the effectiveness of various algorithms in detecting momentary depressive feelings. These findings, consistent with previous studies, highlight the superiority of transfer learning algorithms in NLP tasks, particularly in the detection of momentary depressive feelings in social media data. This study used state-of-the-art algorithms that leverage large amounts of data and generalize to new tasks. Transfer learning algorithms are especially adept at processing large data sets and can identify patterns and features that are difficult to capture with traditional ML algorithms.

Within the domain of binary and multilabel classification, our analysis uncovers a compelling revelation. Specifically, our findings highlight the intricate nature of posts conveying depressive emotions, irrespective of their varying degrees or gradations. These posts present a formidable obstacle for classification models due to their nuanced character, encompassing a broad range of emotional states that defy simple categorization. This insight serves as a poignant reminder of the challenges inherent in accurately classifying such nuanced sentiments, especially when addressing the realm of depressive expressions.

The significant findings of this study have the potential to make a meaningful impact on mental health, particularly in momentary depressive feelings detection. Early detection of momentary depressive feelings can pave the way for timely interventions and ultimately improve mental health outcomes. The approach presented in this study could be integrated into social media monitoring tools to identify individuals who are at risk of developing depression or who may benefit from mental health interventions. This approach could lead to more efficient and effective mental health interventions, resulting in better outcomes for individuals with mental health conditions and reducing the burden imposed by mental health disorders. The findings of this study provide a beneficial stepping stone for the development of new and innovative approaches to mental health monitoring and intervention.

Although the findings of the study are promising, there are limitations. First, the study only focused on momentary depressive feelings, and the results may not be generalizable to other mental health conditions such as anxiety, stress, or other mood disorders. Second, the study relied solely on X data, which may not represent the broader population. Future research could investigate the use of other social media platforms or clinical data to assess the effectiveness of contextual language approaches for detecting mental health conditions in a more diverse population. Additionally, these approaches may be limited in their ability to capture the nuances and complexity of mental health conditions. Although this method is effective, it may not always detect posts that express subtle or indirect signs of mental health conditions. Future studies could explore the use of ML techniques to learn from the data to detect momentary depressive feelings rather than relying on predefined lexicons. This approach could lead to more accurate results and more effective detection of mental health conditions. Finally, the study only used English posts, which further limits the generalizability of the findings.

Conclusions

This study aimed to detect momentary depressive feelings using X data and contextual language approaches. In this study, we applied a methodology consisting of data collection, manual labeling, and post analysis with contextual language approaches. A lexicon containing 32 keywords relevant to depressive feelings was established, and then, using this lexicon, *Twint* was used to extract posts from January 2022 to December 2022. Six baseline algorithms were used for the detection of momentary depressive feelings, and the results were evaluated using AUC, accuracy, sensitivity, specificity, precision, and F_1 -score.

Our results showed that DistilBERT, a transfer learning algorithm, had the highest performance in terms of the evaluation metrics described. The study found that transfer learning algorithms are promising tools in NLP tasks, for example, extracting knowledge and detecting patterns in posts, particularly in the detection of momentary depressive feelings.

Our findings demonstrated X data can be used for the detection of momentary depressive feelings. This is achieved through the development of an automated framework for continuously monitoring and detecting individuals' real-time mental states. These findings have significant implications for timely mental health interventions. Early detection of momentary depressive feelings can prevent the escalation of these feelings to more severe depressive symptoms and reduce the burden imposed on people and society. This methodology can be easily applied to large X data sets, making it a useful tool for monitoring depressive symptoms on a large scale. Moreover, this methodology can be improved to be applied to other social media platforms and various mental health conditions. Overall, this study contributes to the growing body of research on using social media data for mental health research. Our approach provides a useful tool for researchers interested in studying momentary depressive feelings using social media data.

Acknowledgments

This work was supported by Refresh Inc and Mitacs through its Accelerate program (grant IT27060).

Data Availability

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

Authors' Contributions

All authors conceptualized the idea and wrote and reviewed the paper. AAJ designed and implemented the experiment; collected, curated, and analyzed the data; and prepared the figures and visualizations. RJS looked after the administration, supervised the study, and provided funding and resources.

Conflicts of Interest

None declared.

Multimedia Appendix 1

Evaluation performance for multilabel classification.

[[DOCX File, 553 KB - ai_v2i1e49531_app1.docx](#)]

References

1. Depression and other common mental disorders: global health estimates. World Health Organization. 2017. URL: <https://www.who.int/publications/i/item/depression-global-health-estimates> [accessed 2023-11-02]
2. Kondo N, Kazama M, Suzuki K, Yamagata Z. Impact of mental health on daily living activities of Japanese elderly. *Prev Med* 2008;46(5):457-462. [doi: [10.1016/j.ypmed.2007.12.007](https://doi.org/10.1016/j.ypmed.2007.12.007)] [Medline: [18258290](https://pubmed.ncbi.nlm.nih.gov/18258290/)]
3. He R, Wei J, Huang K, Yang H, Chen Y, Liu Z, et al. Nonpharmacological interventions for subthreshold depression in adults: a systematic review and network meta-analysis. *Psychiatry Res* 2022;317:114897. [doi: [10.1016/j.psychres.2022.114897](https://doi.org/10.1016/j.psychres.2022.114897)] [Medline: [36242840](https://pubmed.ncbi.nlm.nih.gov/36242840/)]
4. Ashaie SA, Hurwitz R, Cherney LR. Depression and subthreshold depression in stroke-related aphasia. *Arch Phys Med Rehabil* 2019;100(7):1294-1299. [doi: [10.1016/j.apmr.2019.01.024](https://doi.org/10.1016/j.apmr.2019.01.024)] [Medline: [30831094](https://pubmed.ncbi.nlm.nih.gov/30831094/)]
5. Jeon HJ, Park JI, Fava M, Mischoulon D, Sohn JH, Seong S, et al. Feelings of worthlessness, traumatic experience, and their comorbidity in relation to lifetime suicide attempt in community adults with major depressive disorder. *J Affect Disord* 2014;166:206-212. [doi: [10.1016/j.jad.2014.05.010](https://doi.org/10.1016/j.jad.2014.05.010)] [Medline: [25012433](https://pubmed.ncbi.nlm.nih.gov/25012433/)]
6. Aldao A, Nolen-Hoeksema S, Schweizer S. Emotion-regulation strategies across psychopathology: a meta-analytic review. *Clin Psychol Rev* 2010;30(2):217-237. [doi: [10.1016/j.cpr.2009.11.004](https://doi.org/10.1016/j.cpr.2009.11.004)] [Medline: [20015584](https://pubmed.ncbi.nlm.nih.gov/20015584/)]
7. An JH, Jeon HJ, Cho SJ, Chang SM, Kim BS, Hahm BJ, et al. Subthreshold lifetime depression and anxiety are associated with increased lifetime suicide attempts: a Korean nationwide study. *J Affect Disord* 2022;302:170-176. [doi: [10.1016/j.jad.2022.01.046](https://doi.org/10.1016/j.jad.2022.01.046)] [Medline: [35038481](https://pubmed.ncbi.nlm.nih.gov/35038481/)]
8. Schnyer RN, Allen JB. Chapter 2—Depression defined: symptoms, epidemiology, etiology, treatment. In: *Acupuncture in the Treatment of Depression: A Manual for Practice and Research*. Burlington, MA: Churchill Livingstone; 2001:8-30.
9. Noyes BK, Munoz DP, Khalid-Khan S, Brietzke E, Booi L. Is subthreshold depression in adolescence clinically relevant? *J Affect Disord* 2022;309:123-130. [doi: [10.1016/j.jad.2022.04.067](https://doi.org/10.1016/j.jad.2022.04.067)] [Medline: [35429521](https://pubmed.ncbi.nlm.nih.gov/35429521/)]
10. Mitchell LM, Joshi U, Patel V, Lu C, Naslund JA. Economic evaluations of internet-based psychological interventions for anxiety disorders and depression: a systematic review. *J Affect Disord* 2021;284:157-182 [FREE Full text] [doi: [10.1016/j.jad.2021.01.092](https://doi.org/10.1016/j.jad.2021.01.092)] [Medline: [33601245](https://pubmed.ncbi.nlm.nih.gov/33601245/)]
11. Perestelo-Perez L, Barraca J, Peñate W, Rivero-Santana A, Alvarez-Perez Y. Mindfulness-based interventions for the treatment of depressive rumination: systematic review and meta-analysis. *Int J Clin Health Psychol* 2017;17(3):282-295 [FREE Full text] [doi: [10.1016/j.ijchp.2017.07.004](https://doi.org/10.1016/j.ijchp.2017.07.004)] [Medline: [30487903](https://pubmed.ncbi.nlm.nih.gov/30487903/)]
12. Figuerêdo JSL, Maia ALLM, Calumby RT. Early depression detection in social media based on deep learning and underlying emotions. *Online Soc Netw Media* 2022;31:100225. [doi: [10.1016/j.osnem.2022.100225](https://doi.org/10.1016/j.osnem.2022.100225)]
13. Cacheda F, Fernandez D, Novoa FJ, Carneiro V. Early detection of depression: social network analysis and random forest techniques. *J Med Internet Res* 2019;21(6):e12554 [FREE Full text] [doi: [10.2196/12554](https://doi.org/10.2196/12554)] [Medline: [31199323](https://pubmed.ncbi.nlm.nih.gov/31199323/)]
14. Vahia VN. Diagnostic and statistical manual of mental disorders 5: a quick glance. *Indian J Psychiatry* 2013;55(3):220-223 [FREE Full text] [doi: [10.4103/0019-5545.117131](https://doi.org/10.4103/0019-5545.117131)] [Medline: [24082241](https://pubmed.ncbi.nlm.nih.gov/24082241/)]
15. Kroenke K, Spitzer RL. The PHQ-9: a new depression diagnostic and severity measure. *Psychiatr Ann* 2002;32(9):509-515. [doi: [10.3928/0048-5713-20020901-06](https://doi.org/10.3928/0048-5713-20020901-06)]

16. Ahmed T, Qassem M, Kyriacou PA. Physiological monitoring of stress and major depression: a review of the current monitoring techniques and considerations for the future. *Biomed Signal Process Control* 2022;75:103591. [doi: [10.1016/j.bspc.2022.103591](https://doi.org/10.1016/j.bspc.2022.103591)]
17. Savekar A, Tarai S, Singh M. Linguistic markers in individuals with symptoms of depression in bi-multilingual context. In: Paul S, Bhattacharya P, Bit A, editors. *Early Detection of Neurological Disorders Using Machine Learning Systems*. Hershey, PA: IGI Global; 2019:216-240.
18. Tølbøll KB. Linguistic features in depression: a meta-analysis. *J Lang Works* 2019;4(2):39-59 [FREE Full text]
19. Smirnova D, Cumming P, Sloeva E, Kuvshinova N, Romanov D, Nosachev G. Language patterns discriminate mild depression from normal sadness and euthymic state. *Front Psychiatry* 2018;9:105 [FREE Full text] [doi: [10.3389/fpsy.2018.00105](https://doi.org/10.3389/fpsy.2018.00105)] [Medline: [29692740](https://pubmed.ncbi.nlm.nih.gov/29692740/)]
20. Hutto C, Gilbert E. VADER: a parsimonious rule-based model for sentiment analysis of social media text. 2014 Presented at: Proceedings of the International AAAI Conference on Web and Social Media; May 16, 2014; Ann Arbor, MI p. 216-225 URL: <https://ojs.aaai.org/index.php/ICWSM/article/view/14550> [doi: [10.1609/icwsm.v8i1.14550](https://doi.org/10.1609/icwsm.v8i1.14550)]
21. Islam MR, Kabir MA, Ahmed A, Kamal ARM, Wang H, Ulhaq A. Depression detection from social network data using machine learning techniques. *Health Inf Sci Syst* 2018;6(1):8 [FREE Full text] [doi: [10.1007/s13755-018-0046-0](https://doi.org/10.1007/s13755-018-0046-0)] [Medline: [30186594](https://pubmed.ncbi.nlm.nih.gov/30186594/)]
22. Mikolov T, Chen K, Corrado G, Dean J. Efficient estimation of word representations in vector space. arXiv Preprint posted online on January 16, 2013. [FREE Full text] [doi: [10.48550/arXiv.1301.3781](https://doi.org/10.48550/arXiv.1301.3781)]
23. Hassan A, Mahmood A. Deep learning approach for sentiment analysis of short texts. : IEEE; 2017 Presented at: 2017 3rd International Conference on Control, Automation and Robotics (ICCAR); April 24-26, 2017; Nagoya, Japan. [doi: [10.1109/iccar.2017.7942788](https://doi.org/10.1109/iccar.2017.7942788)]
24. Behera RK, Jena M, Rath SK, Misra S. Co-LSTM: convolutional LSTM model for sentiment analysis in social big data. *Inf Process Manag* 2021;58(1):102435. [doi: [10.1016/j.ipm.2020.102435](https://doi.org/10.1016/j.ipm.2020.102435)]
25. 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] [doi: [10.18653/v1/N19-1423](https://doi.org/10.18653/v1/N19-1423)]
26. Zhu C. Chapter 2—The basics of natural language processing. In: *Machine Reading Comprehension: Algorithms and Practice*. Amsterdam: Elsevier; 2021:27-46.
27. Steinert S, Dennis MJ. Emotions and digital well-being: on social media's emotional affordances. *Philos Technol* 2022;35(2):36 [FREE Full text] [doi: [10.1007/s13347-022-00530-6](https://doi.org/10.1007/s13347-022-00530-6)] [Medline: [35450167](https://pubmed.ncbi.nlm.nih.gov/35450167/)]
28. Van Le D, Montgomery J, Kirkby KC, Scanlan J. Risk prediction using natural language processing of electronic mental health records in an inpatient forensic psychiatry setting. *J Biomed Inform* 2018;86:49-58 [FREE Full text] [doi: [10.1016/j.jbi.2018.08.007](https://doi.org/10.1016/j.jbi.2018.08.007)] [Medline: [30118855](https://pubmed.ncbi.nlm.nih.gov/30118855/)]
29. Patel R, Lloyd T, Jackson R, Ball M, Shetty H, Broadbent M, et al. Mood instability and clinical outcomes in mental health disorders: a natural language processing (NLP) study. *Eur Psychiatr* 2016;33(S1):s224-s224. [doi: [10.1016/j.eurpsy.2016.01.551](https://doi.org/10.1016/j.eurpsy.2016.01.551)]
30. Negriff S, Lynch FL, Cronkite DJ, Pardee RE, Penfold RB. Using natural language processing to identify child maltreatment in health systems. *Child Abuse Negl* 2023;138:106090. [doi: [10.1016/j.chiabu.2023.106090](https://doi.org/10.1016/j.chiabu.2023.106090)] [Medline: [36758373](https://pubmed.ncbi.nlm.nih.gov/36758373/)]
31. Yusof N, Lin C, He Y. Sentiment analysis in social media. In: Alhajj R, Rokne J, editors. *Encyclopedia of Social Network Analysis and Mining*. New York: Springer; 2018:1-13.
32. Wang X, Zhang C, Ji Y, Sun L, Wu L, Bao Z. A depression detection model based on sentiment analysis in micro-blog social network. In: *Trends and Applications in Knowledge Discovery and Data Mining*. Berlin, Heidelberg: Springer; 2013:201-213.
33. De Choudhury M, Gamon M, Counts S, Horvitz E. Predicting depression via social media. 2021 Presented at: Proceedings of the International AAAI Conference on Web and Social Media; August 3, 2021; Online p. 128-137 URL: <https://ojs.aaai.org/index.php/ICWSM/article/view/14432> [doi: [10.1609/icwsm.v7i1.14432](https://doi.org/10.1609/icwsm.v7i1.14432)]
34. William D, Suhartono D. Text-based depression detection on social media posts: a systematic literature review. *Procedia Comput Sci* 2021;179:582-589 [FREE Full text] [doi: [10.1016/j.procs.2021.01.043](https://doi.org/10.1016/j.procs.2021.01.043)]
35. Liu T, Meyerhoff J, Eichstaedt JC, Karr CJ, Kaiser SM, Kording KP, et al. The relationship between text message sentiment and self-reported depression. *J Affect Disord* 2022;302:7-14 [FREE Full text] [doi: [10.1016/j.jad.2021.12.048](https://doi.org/10.1016/j.jad.2021.12.048)] [Medline: [34963643](https://pubmed.ncbi.nlm.nih.gov/34963643/)]
36. Neuman Y, Cohen Y, Assaf D, Kedma G. Proactive screening for depression through metaphorical and automatic text analysis. *Artif Intell Med* 2012;56(1):19-25. [doi: [10.1016/j.artmed.2012.06.001](https://doi.org/10.1016/j.artmed.2012.06.001)] [Medline: [22771201](https://pubmed.ncbi.nlm.nih.gov/22771201/)]
37. Bacsu JD, O'Connell ME, Cammer A, Azizi M, Grewal K, Poole L, et al. Using Twitter to understand the COVID-19 experiences of people with dementia: infodemiology study. *J Med Internet Res* 2021 Mar 03;23(2):e26254 [FREE Full text] [doi: [10.2196/26254](https://doi.org/10.2196/26254)] [Medline: [33468449](https://pubmed.ncbi.nlm.nih.gov/33468449/)]
38. Orabi AH, Buddhitha P, Orabi MH, Inkpen D. Deep learning for depression detection of Twitter users. 2018 Presented at: Proceedings of the Fifth Workshop on Computational Linguistics and Clinical Psychology: From Keyboard to Clinic; June 5, 2018; New Orleans, LA p. 88-97. [doi: [10.18653/v1/w18-0609](https://doi.org/10.18653/v1/w18-0609)]

39. Oscar N, Fox PA, Croucher R, Wernick R, Keune J, Hooker K. Machine learning, sentiment analysis, and tweets: an examination of Alzheimer's disease stigma on Twitter. *J Gerontol B Psychol Sci Soc Sci* 2017;72(5):742-751 [[FREE Full text](#)] [doi: [10.1093/geronb/gbx014](https://doi.org/10.1093/geronb/gbx014)] [Medline: [28329835](https://pubmed.ncbi.nlm.nih.gov/28329835/)]
40. McManus K, Mallory EK, Goldfeder RL, Haynes WA, Tatum JD. Mining Twitter data to improve detection of schizophrenia. *AMIA Jt Summits Transl Sci Proc* 2015;2015:122-126 [[FREE Full text](#)] [Medline: [26306253](https://pubmed.ncbi.nlm.nih.gov/26306253/)]
41. McCosker A, Gerrard Y. Hashtagging depression on Instagram: towards a more inclusive mental health research methodology. *New Media Soc* 2021;23(7):1899-1919. [doi: [10.1177/1461444820921349](https://doi.org/10.1177/1461444820921349)]
42. Cha J, Kim S, Park E. A lexicon-based approach to examine depression detection in social media: the case of Twitter and university community. *Humanit Soc Sci Commun* 2022;9(1):325 [[FREE Full text](#)] [doi: [10.1057/s41599-022-01313-2](https://doi.org/10.1057/s41599-022-01313-2)] [Medline: [36159708](https://pubmed.ncbi.nlm.nih.gov/36159708/)]
43. O'Connor C, Joffe H. Intercoder reliability in qualitative research: debates and practical guidelines. *Int J Qual Methods* 2020;19:1-13 [[FREE Full text](#)] [doi: [10.1177/1609406919899220](https://doi.org/10.1177/1609406919899220)]
44. Lan Z, Chen M, Goodman S, Gimpel K, Sharma P, Soricut R. Albert: A Lite BERT for self-supervised learning of language representations. arXiv Preprint posted online on September 26, 2019. [[FREE Full text](#)] [doi: [10.48550/arXiv.1909.11942](https://doi.org/10.48550/arXiv.1909.11942)]
45. 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. [[FREE Full text](#)] [doi: [10.48550/arXiv.1907.11692](https://doi.org/10.48550/arXiv.1907.11692)]
46. 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. [[FREE Full text](#)]
47. Raitoharju J. Chapter 3—Convolutional neural networks. In: Iosifidis A, Tefas A, editors. *Deep Learning for Robot Perception and Cognition*. Amsterdam: Academic Press; 2022:35-69.
48. Umer M, Sadiq S, Karamti H, Abdulmajid Eshawi A, Nappi M, Usman Sana M, et al. ETCNN: extra tree and convolutional neural network-based ensemble model for COVID-19 tweets sentiment classification. *Pattern Recognit Lett* 2022;164:224-231 [[FREE Full text](#)] [doi: [10.1016/j.patrec.2022.11.012](https://doi.org/10.1016/j.patrec.2022.11.012)] [Medline: [36407854](https://pubmed.ncbi.nlm.nih.gov/36407854/)]
49. Aslan S. A deep learning-based sentiment analysis approach (MF-CNN-BILSTM) and topic modeling of tweets related to the Ukraine—Russia conflict. *Appl Soft Comput* 2023;143:110404. [doi: [10.1016/j.asoc.2023.110404](https://doi.org/10.1016/j.asoc.2023.110404)]
50. Sangeetha J, Kumaran U. A hybrid optimization algorithm using BiLSTM structure for sentiment analysis. *Meas Sens* 2023;25:100619 [[FREE Full text](#)] [doi: [10.1016/j.measen.2022.100619](https://doi.org/10.1016/j.measen.2022.100619)]
51. Ursenbach J, O'Connell ME, Neiser J, Tierney MC, Morgan D, Kosteniuk J, et al. Scoring algorithms for a computer-based cognitive screening tool: an illustrative example of overfitting machine learning approaches and the impact on estimates of classification accuracy. *Psychol Assess* 2019;31(11):1377-1382. [doi: [10.1037/pas0000764](https://doi.org/10.1037/pas0000764)] [Medline: [31414853](https://pubmed.ncbi.nlm.nih.gov/31414853/)]
52. 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;212(1):38-43. [doi: [10.2214/AJR.18.20224](https://doi.org/10.2214/AJR.18.20224)] [Medline: [30332290](https://pubmed.ncbi.nlm.nih.gov/30332290/)]
53. Makita M, Mas-Bleda A, Morris S, Thelwall M. Mental health discourses on Twitter during mental health awareness week. *Issues Ment Health Nurs* 2021;42(5):437-450. [doi: [10.1080/01612840.2020.1814914](https://doi.org/10.1080/01612840.2020.1814914)] [Medline: [32926796](https://pubmed.ncbi.nlm.nih.gov/32926796/)]
54. Williams ML, Burnap P, Sloan L. Towards an ethical framework for publishing Twitter data in social research: taking into account users' views, online context and algorithmic estimation. *Sociology* 2017;51(6):1149-1168 [[FREE Full text](#)] [doi: [10.1177/0038038517708140](https://doi.org/10.1177/0038038517708140)] [Medline: [29276313](https://pubmed.ncbi.nlm.nih.gov/29276313/)]
55. Roy A, Ojha M. Twitter sentiment analysis using deep learning models. : IEEE; 2020 Presented at: 2020 IEEE 17th India Council International Conference (INDICON); December 10-13, 2020; New Delhi, India p. 1-6. [doi: [10.1109/indicon49873.2020.9342279](https://doi.org/10.1109/indicon49873.2020.9342279)]

Abbreviations

- ALBERT:** A Lite BERT
AUC: area under the curve
BERT: Bidirectional Encoder Representations From Transformers
BiLSTM: bidirectional long short-term memory
CNN: convolutional neural network
CV: cross-validation
DistilBERT: Distilled BERT
ML: machine learning
NLP: natural language processing
RoBERTa: Robustly Optimized BERT Approach
VADER: Valence Aware Dictionary for Sentiment Reasoning

Edited by C Xiao; submitted 31.05.23; peer-reviewed by F Rudzicz, S Matsuda; comments to author 20.08.23; revised version received 06.09.23; accepted 27.10.23; published 27.11.23.

Please cite as:

Jamali AA, Berger C, Spiteri RJ

Momentary Depressive Feeling Detection Using X (Formerly Twitter) Data: Contextual Language Approach

JMIR AI 2023;2:e49531

URL: <https://ai.jmir.org/2023/1/e49531>

doi: [10.2196/49531](https://doi.org/10.2196/49531)

PMID: [38875532](https://pubmed.ncbi.nlm.nih.gov/38875532/)

©Ali Akbar Jamali, Corinne Berger, Raymond J Spiteri. Originally published in JMIR AI (<https://ai.jmir.org>), 27.11.2023. This is an open-access article distributed under the terms of the Creative Commons Attribution License (<https://creativecommons.org/licenses/by/4.0/>), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR AI, is properly cited. The complete bibliographic information, a link to the original publication on <https://www.ai.jmir.org/>, as well as this copyright and license information must be included.

Original Paper

Developing Ethics and Equity Principles, Terms, and Engagement Tools to Advance Health Equity and Researcher Diversity in AI and Machine Learning: Modified Delphi Approach

Rachele Hendricks-Sturup¹, MA, MS, DHSc; Malaika Simmons¹, MSHE; Shilo Anders², PhD; Kammarauche Aneni³, MBBS, MHS; Ellen Wright Clayton², MD, JD; Joseph Coco², MS; Benjamin Collins², MD, MS, MA; Elizabeth Heitman⁴, PhD; Sajid Hussain⁵, PhD; Karuna Joshi⁶, PhD; Josh Lemieux⁷; Laurie Lovett Novak², PhD; Daniel J Rubin⁸, MD, MSc; Anil Shanker⁹, PhD; Talitha Washington¹⁰, PhD; Gabriella Waters¹¹; Joyce Webb Harris², MA; Rui Yin¹², PhD; Teresa Wagner¹³, MS, DrPH; Zhijun Yin², MS, PhD; Bradley Malin², PhD

¹National Alliance Against Disparities in Patient Health, Woodbridge, VA, United States

²Vanderbilt University Medical Center, Nashville, TN, United States

³Yale University, New Haven, CT, United States

⁴University of Texas Southwestern Medical Center, Dallas, TX, United States

⁵Fisk University, Nashville, TN, United States

⁶University of Maryland, Baltimore County, Baltimore, MD, United States

⁷OCHIN, Portland, OR, United States

⁸Temple University, Philadelphia, PA, United States

⁹Meharry Medical College, Nashville, TN, United States

¹⁰AUC Data Science Initiative, Clark Atlanta University, Atlanta, GA, United States

¹¹Morgan State University, Center for Equitable AI & Machine Learning Systems, Baltimore, MD, United States

¹²University of Florida, Gainesville, FL, United States

¹³University of North Texas Health Science Center, SaferCare Texas, Fort Worth, TX, United States

Corresponding Author:

Rachele Hendricks-Sturup, MA, MS, DHSc
National Alliance Against Disparities in Patient Health
2700 Neabsco Common Place
Suite 101
Woodbridge, VA, 22191
United States
Phone: 1 (571) 316 5116
Email: hendricks-sturup@nadph.org

Abstract

Background: Artificial intelligence (AI) and machine learning (ML) technology design and development continues to be rapid, despite major limitations in its current form as a practice and discipline to address all sociohumanitarian issues and complexities. From these limitations emerges an imperative to strengthen AI and ML literacy in underserved communities and build a more diverse AI and ML design and development workforce engaged in health research.

Objective: AI and ML has the potential to account for and assess a variety of factors that contribute to health and disease and to improve prevention, diagnosis, and therapy. Here, we describe recent activities within the Artificial Intelligence/Machine Learning Consortium to Advance Health Equity and Researcher Diversity (AIM-AHEAD) Ethics and Equity Workgroup (EEWG) that led to the development of deliverables that will help put ethics and fairness at the forefront of AI and ML applications to build equity in biomedical research, education, and health care.

Methods: The AIM-AHEAD EEWG was created in 2021 with 3 cochairs and 51 members in year 1 and 2 cochairs and ~40 members in year 2. Members in both years included AIM-AHEAD principal investigators, coinvestigators, leadership fellows, and research fellows. The EEWG used a modified Delphi approach using polling, ranking, and other exercises to facilitate

discussions around tangible steps, key terms, and definitions needed to ensure that ethics and fairness are at the forefront of AI and ML applications to build equity in biomedical research, education, and health care.

Results: The EEWG developed a set of ethics and equity principles, a glossary, and an interview guide. The ethics and equity principles comprise 5 core principles, each with subparts, which articulate best practices for working with stakeholders from historically and presently underrepresented communities. The glossary contains 12 terms and definitions, with particular emphasis on optimal development, refinement, and implementation of AI and ML in health equity research. To accompany the glossary, the EEWG developed a concept relationship diagram that describes the logical flow of and relationship between the definitional concepts. Lastly, the interview guide provides questions that can be used or adapted to garner stakeholder and community perspectives on the principles and glossary.

Conclusions: Ongoing engagement is needed around our principles and glossary to identify and predict potential limitations in their uses in AI and ML research settings, especially for institutions with limited resources. This requires time, careful consideration, and honest discussions around what classifies an engagement incentive as meaningful to support and sustain their full engagement. By slowing down to meet historically and presently underresourced institutions and communities where they are and where they are capable of engaging and competing, there is higher potential to achieve needed diversity, ethics, and equity in AI and ML implementation in health research.

(*JMIR AI 2023;2:e52888*) doi:[10.2196/52888](https://doi.org/10.2196/52888)

KEYWORDS

artificial intelligence; AI; Delphi; disparities; disparity; engagement; equitable; equities; equity; ethic; ethical; ethics; fair; fairness; health disparities; health equity; humanitarian; machine learning; ML

Introduction

Recent events and academic literature have underscored a role for the field of artificial intelligence (AI) and machine learning (ML) technology to take all stakeholders' impressions and concerns into account to inform approaches for achieving health equity [1-5]. It has also become imperative to strengthen AI and ML literacy in underserved communities and build a more diverse workforce in AI and ML design and development. However, whether as a practice or as an academic discipline, AI and ML are not yet engineered to address all sociohumanitarian issues and complexities. This is especially true for socially and economically marginalized communities whose members are frequently unheard or have limited engagement in research, discovery, and innovation pipelines for cultivating shared prosperity.

The general population still has limited knowledge about AI and ML, with 1 study reporting that only about one-quarter of people have heard of AI or ML, and only about half are at least somewhat aware of AI and ML [6]. Furthermore, individuals and communities who are subject to potentially detrimental outcomes (persons with mental health care needs and disabilities, persons with marginalized racial or ethnic identities, etc) may be more aware of the potential harms of AI and ML, particularly when it comes to the risk of harm from bias [7,8]. Thus, people who are presently or historically underserved or marginalized may be particularly concerned that they will be harmed by AI or ML technologies, especially in cases where AI or ML is used or applied without their awareness.

The overall lack of understanding about AI and ML and the awareness of bias among historically and presently marginalized

populations could result in limited trust in the technology and its use. To build trust among those most subject to bias or at risk of detrimental outcomes, it is critical for AI and ML developers to assess their own reliability and adapt their practices to build trustworthiness with the most vulnerable stakeholders. In this context, it is also important to recognize that trust varies across and within populations, and people may have more or less trust in health care technologies based on factors such as previous experience of racial bias [9].

If implemented responsibly, AI and ML has the power to account for and assess a variety of factors that contribute to health and disease to improve prevention, diagnosis, and therapy. The ability to predict the risk of adverse health outcomes and identify high-risk patients for targeted preventive interventions offers tremendous potential to improve the health of individuals and medically underserved populations [10,11].

A great deal of AI and ML today is developed without meaningful engagement of individuals and communities, even when those individuals and communities have (knowingly or unknowingly) generated data used by AI and ML models. When there are proactive efforts to engage communities in AI and ML design, development, or application, various factors may negatively affect how people respond (Textbox 1). For instance, failure to educate about AI and ML and contextualize its impact on an individual and their community may bias individuals' consent to contribute data to build such technologies and, subsequently, lead to biased outcomes in terms of who benefits from the technology's development and application. Consequently, poor engagement can exacerbate inequities in the creation, development, and application of AI and ML.

Textbox 1. Factors that may engender inequitable access to artificial intelligence (AI) and machine learning (ML) or demotivate participation in AI and ML.

Factors that demotivate participation in AI and ML

- Cultural norms or expectations that discourage the use of AI and ML technology
- Fear and reservations that the AI and ML tool may be used to cause harm
- The history of major AI and ML–developing institutions is not inclusive of all communities, thus defying communities’ trust
- The lack of access to high-performance infrastructure and resources needed to execute AI and ML models
- The lack of interest, excitement, or perception of “hype”
- Unaddressed confusion, misinformation, or disillusionment

Factors that exacerbate inequitable access to the benefit of AI and ML

- Asymmetric ability to extract value from AI and ML
- Insufficient access to the internet, data, and data services (ie, digital divide)
- Insufficient funding or economic opportunities
- There is an intractable disagreement and power imbalance between stakeholders about how AI and ML should be used or applied
- Lack of institutional leadership or commitment
- Limited experience, knowledge, and education
- Sociocultural factors affecting digital access and inclusion

The underengagement of communities in research, development, and use of AI and ML often reflects limited knowledge and crucial misunderstandings about AI and ML, including how it is used in health care settings to advance health-related innovations and solutions. Thus, stronger, more targeted, and more intentional engagement is required to help these groups identify and address real or potential harms associated with the problematic implementation of AI and ML in high-consequence settings. To address this challenge, the US National Institutes of Health’s Artificial Intelligence/Machine Learning Consortium to Advance Health Equity and Researcher Diversity (AIM-AHEAD) was established in 2021 with a mission to address factors that undermine achieving health equity through the design, use, and application of AI and ML, including the lack of the following:

- An adequately diverse workforce
- Adequate data and data infrastructure
- Adequate community engagement
- Adequate oversight, governance, and accountability
- Consensus that ethics can strengthen innovation

The tension between individual desires and population needs challenges ethics and equity in AI and ML settings. Thus, the Ethics and Equity Workgroup (EEWG) was formed within the AIM-AHEAD Consortium to ensure that ethics and fairness are at the forefront of AI and ML applications to build equity in biomedical research, education, and health care. Activities within the workgroup have included deliberations and

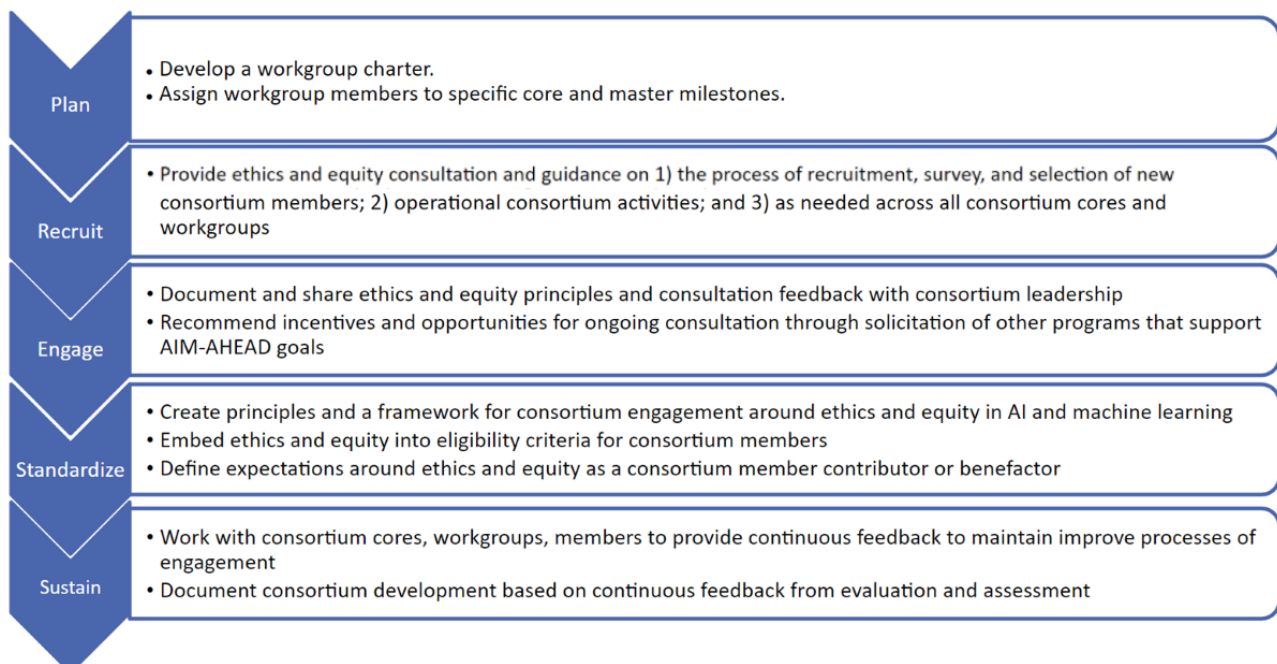
discussions to develop and reach consensus on actionable guiding principles, a glossary of key terms, and other engagement tools to encourage greater attention to ethics and equity in AI and ML development. This study describes these activities with the intent to serve and inform the AIM-AHEAD community of stakeholders; external consortia, organizations, and communities that have goals similar to the AIM-AHEAD; and those interested in ethical and equitable AI and ML development and applications more broadly.

Methods

Workgroup Establishment

The AIM-AHEAD EEWG was created in 2021 to guide the ethical and equitable development and implementation of AI and ML tools and processes broadly within the AIM-AHEAD. Simultaneously, an Equitable Policy Development Workgroup was developed within the AIM-AHEAD Infrastructure Core. To ensure rapid and coordinated progress with respect to embedding ethics and equity into AIM-AHEAD activities, both within and outside of the Infrastructure Core, the EEWG’s efforts were harmonized and merged with the Infrastructure Core’s Equitable Policy Development Workgroup upon recommendation by the EEWG cochair and multiple principal investigators for the AIM-AHEAD Infrastructure Core. The newly reconfigured EEWG began by defining its scope of activities (Figure 1).

Figure 1. Artificial Intelligence/Machine Learning Consortium to Advance Health Equity and Researcher Diversity (AIM-AHEAD) Ethics and Equity Workgroup's scope of activities. AI: artificial intelligence.



Workgroup Membership

At the start of the program in year 1, the EEWG was comprised of 51 members (AIM-AHEAD principal investigators and coinvestigators) and 3 cochairs. AIM-AHEAD participants either requested to join or were selected to join by their project leaders within the program. During year 2, the EEWG's membership was consolidated into 2 cochairs and approximately 40 AIM-AHEAD principal investigators, coinvestigators, leadership fellows, and research fellows. This reduction in EEWG cochairs and members occurred for two main reasons: (1) time and effort among members were reallocated to other activities within the AIM-AHEAD (administrative planning for regional hubs, research, etc), and (2) given the evolution of the program over time, the year 1 members were provided an opportunity to recommit to the EEWG for year 2. In both years, EEWG cochairs and members represented a variety of academic disciplines and focus areas, including but not limited to medicine, computational science, population health, health science, data science, bioethics, law, community engagement, human-centered design, health disparities research, biological science, social science, and engineering.

Development of a Set of Ethical Principles for AI and ML

The initial effort of the EEWG during year 1 was to produce a set of principles and a glossary to inform the practice of ethics and equity in AI and ML development and implementation in health research. During year 1, members convened in weekly meetings that led to consensus on the development of specific workgroup deliverables. EEWG members reviewed the literature to identify relevant sources with perspectives on ethics, equity, and social determinants of health, especially those that were community driven, and lessons that could inform the development and use of AI and ML in health disparity and disease prevention research [12-27].

To develop the principles, the EEWG used a modified Delphi approach to facilitate discussions around tangible steps that the Consortium should take to ensure that ethics and fairness are at the forefront of AI and ML applications to build equity in biomedical research, education, and health care [28]. Specifically, the EEWG engaged in weekly (year 1) and biweekly (year 2) meetings to suggest, review, and deliberate a corpus of published content and literature considered useful toward integrating ethics and equity into AI and ML development and contributed original thought leadership and content in reaction to the content and literature reviewed to devise actionable principles. The EEWG approached the development of the principles with optimism about the potential of AI and ML to address health disparities by empowering communities, yet with recognition of complex societal challenges: inadequate or misrepresentation in data sets, algorithmic bias, imbalances in communities' access to data and information about themselves, misuses of AI and ML tools, and threats to the civil and human rights of individuals and communities who are or may be subject to illegal or pervasive AI and ML surveillance, to name just a few.

Development of a Glossary

To develop the glossary, during year 1, the EEWG began by defining ways in which outputs of AI and ML can (1) fail to be informative or useful for individuals and groups; (2) distinguish among individuals in inappropriate ways as a result of bias, failure of inclusion, or misuse; or (3) be poorly vetted by individuals and groups who are or may be subject to potentially harmful actions and decisions made by key or authoritative stakeholders that rely on AI and ML for decision support as a result of insufficient engagement with key stakeholders, including data participants.

Using a modified Delphi approach that likewise involved polling, ranking, and other exercises, consensus was reached

on terms to define [29]. During its meetings, the EEWG discussed all possible terms that would be key to define to inform the ethical and equitable development and application of AI and ML, followed by 2 rounds of ranking and polling exercises to narrow their suggestions to 12 sentinel terms. Sentinel terms discussed during meetings, for example, included demographical terms such as self-defined or assigned race, ethnicity, sex, ability, and gender that can lead to errors in the development of AI and ML, which can in turn lead to potentially irreversible, intergenerational, and multigenerational harm to individuals and groups subjected to decisions informed by or based on AI and ML outputs. During year 2, remote meetings were held on a biweekly basis to further deliberate and refine the principles and glossary. Refinements were based on expert stakeholder feedback gathered through a survey among participants in the AIM-AHEAD pilot project and during remote convenings.

Development of an Interview Guide

The EEWG initially sought to conduct a quantitative survey to assess how AIM-AHEAD researchers would implement the principles in practice. A draft survey was developed by 2 volunteers within the workgroup, who later shared the draft survey with the broader workgroup for iterative feedback and edits during weekly (year 1) and biweekly (year 2) meetings. The draft survey was also shared with awardees of AIM-AHEAD pilot projects for feedback. As the EEWG deliberated on the feedback, it ultimately determined that a qualitative interview (vs a quantitative survey) would be a more useful approach to garnering AIM-AHEAD researchers' perspectives on implementing the principles in practice. Thereafter, the EEWG met regularly to convert the quantitative survey into an interview guide with the intent of learning the interviewees' perspectives and natural reactions to the AIM-AHEAD ethics and equity principles and glossary.

Ethical Considerations

The EEWG's efforts in developing the interview guide and conducting the interviews were focused exclusively on program-specific planning for the AIM-AHEAD and were not intended as human subjects research. AIM-AHEAD investigators' responses to the interviews were wholly voluntary, and their comments were used exclusively to develop the program's principles and were subject to further assessment for generalizable knowledge.

Results

AIM-AHEAD Ethics and Equity Principles

Overview

Based on the EEWG's internal Delphi process, informed by insights from interviews with AIM-AHEAD investigators, the workgroup articulated 5 core principles, each with subparts, which articulate best practices for working with stakeholders from historically and presently underrepresented communities.

1. Build trust with communities
2. Design and implement AI and ML with intention
3. Cocreate, do not dictate

4. Build capacity
5. Reset the rules

Build Trust With Communities

Researchers should build trust and share power to enable data-driven decision-making among multiple partners—this must be earned through longstanding, sustained relationships in the community, which takes time, investment, and resources to manifest.

- Through authentic community engagement, determine, understand, and deliver value in a manner that is community driven, community defined, and community led.
- Use asset-based language and thinking in collecting, interpreting, and reporting community-level data (in lieu of deficit-based language and thinking).
- Be transparent about the structure of AI models, data that are contextually limited or incomplete, and limitations in the capabilities of data analytics tools and platforms.
- Commit to ongoing engagement and bidirectional communication between AI and ML developers and communities around interventions to address limitations in the capabilities of data analytics tools and platforms.

Design and Implement AI and ML With Intention

Researchers should take collective action and engage in data-driven decision-making toward embedding equity, which requires shared goal setting, design, implementation, and accountability.

- Determine shared goals that serve as a commitment anchor and barometer for cocreated actions.
- Design with intent to overcome root causes of bias to solve or address (vs merely explore) an immediate, ongoing, or systemic problem affecting communities experiencing certain hardships that have contributed to health inequity.
- Develop and implement ongoing AI and ML design mechanisms and procedures to monitor AI and ML algorithms with the goal of preventing or mitigating harm.

Cocreate, Do Not Dictate

Researchers should move from superficial community engagement to true community partnership through meaningful cocreation.

- Develop AI and ML infrastructure, protocols, and programs in partnership with key and affected community stakeholders.
- Avoid tokenizing individuals and communities to achieve asymmetric goals that are or can be perceived as to the detriment of communities.
- Limit the use of computational methods that are or can be perceived as a substitution for data that would be only obtained through strong community engagement.
- Be transparent about the short-, medium-, and long-term sponsorships, investors in, and potential beneficiaries of AI and ML projects.

Build Capacity

Researchers should invest in people, data, and computational technology—today, as community leaders dig into this work,

and tomorrow, as society collectively builds a stronger, more diverse tech talent pipeline.

- Educate stakeholders to enable AI and ML competency across clinical practice, community, and research settings (eg, build AI and ML model fact labels that can summarize or explain algorithms).
- Develop a plan to promote eHealth literacy in marginalized and underserved communities and groups.
- Build equitable access to AI and ML technology, its development, applications, and uses across real-world health contexts including social determinants of health and research.
- Develop a plan for building capacity that includes hiring and supporting a diverse workforce, dedicating funds for sustaining an existing workforce, and creating metrics that allow institutions to measure their success.

Reset the Rules

Researcher should reexamine the mechanisms that hold institutions accountable and resist the urgency of quick fixes to complex issues like systemic racism.

- Engage communities to determine their experiences with and desires to overcome the digital divide and facilitate the equitable inclusion and consideration of populations in AI and ML models and algorithms.
- Create equitable and liberated access to AI and ML development, implementation, and maintenance to oversee and correct model drift and guide entities in their reactions to AI and ML outputs.
- Identify and correct information asymmetries that may lead to communities' lacking pertinent, actionable, and critical information that is exclusively held by powerful institutions.

AIM-AHEAD Ethics and Equity Glossary Terms

Developers of AI and ML platforms and tools must contemplate, anticipate, mitigate, and address potential issues with downstream data aggregation, interpretation, and use. Meeting these goals requires a shared understanding of the terms used in these policies and processes. The EEWG determined that, in many cases, sensitive demographic characteristics (eg, race, ethnicity, sex, ability, and gender) are particularly problematic as variables used in AI and ML because they are often inappropriately understood as being rooted solely or primarily in genetic or phenotypic differences rather than strongly influenced by discriminatory sociohistorical and sociocultural practices.

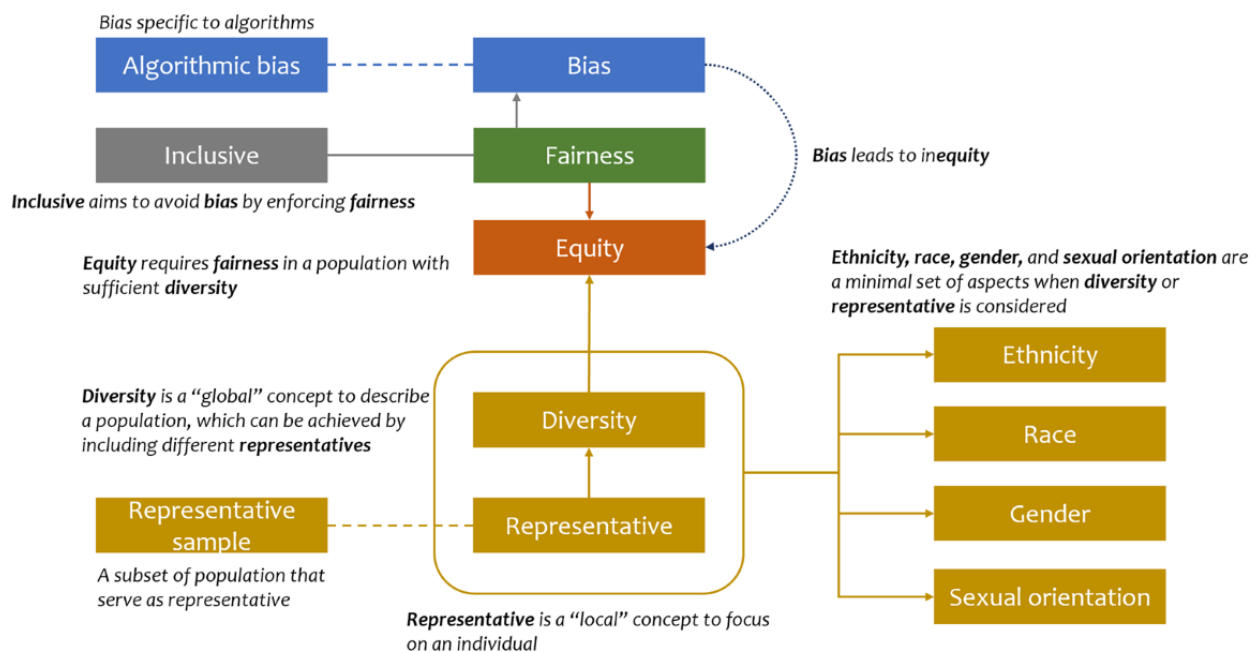
To capture and promote a shared understanding of key terms, the EEWG developed a glossary of 12 words (Table 1) out of 28 considered that follow or build upon existing understandings of these concepts, highlighting their particular importance for the optimal development, refinement, and implementation of AI and ML.

In addition, the EEWG developed a concept relationship diagram that describes the logical flow of and relationship between the definitional concepts described in Table 1 and Figure 2. The center of this diagram is equity, which requires AI developers and implementers to enforce fairness and avoid bias in a population with sufficient diversity by being inclusive. To implement diversity, representatives that are characterized by a minimal set of aspects—ethnicity, race, gender, and sexual orientation—need to be collected. They will form a representative sample if they can reflect the characteristics of a population. A representative sample can mitigate algorithmic bias, which is one specific type of bias.

Table 1. Artificial Intelligence/Machine Learning Consortium to Advance Health Equity and Researcher Diversity (AIM-AHEAD) ethics and equity glossary terms and definitions.

No	Glossary term	AIM-AHEAD definition
1	Ethnicity	Distinct patterns of language, lifestyle, illness, and health beliefs encountered among an individual or representative population, regardless of race, and that may subject the individual or population to bias or discrimination.
2	Race	A social construct or assumption based on patterns in an individual's or representative population's language, lifestyle, and health beliefs and immutable characteristics, such as skin tone, color, or hair texture, regardless of immigration status, socioeconomic status, genetic ancestry, or geographic origin, that may subject the individual's or population to bias, structural racism, or discrimination that would warrant corrective antiracism actions.
3	Bias	Systematic error in information originating, gathering, or assessment activities, leading to selecting or encouraging one outcome or answer over others, which can result in human decisions and values that echo societal or historical inequities and produce inconclusive or limited assumptions about the broader population.
4	Equity	Equity is fairness and justice in policy, practice, and opportunity designed to address the distinct challenges of nondominant social groups with an eye to progressive outcomes. Health equity is the state in which everyone has the opportunity to attain full health potential, and no individual is disadvantaged from achieving this potential because of social position or any other socially defined circumstance.
5	Algorithmic bias	Systematic and repeated errors in the collection and consideration of a variety of factors, including but not limited to the design of the algorithm; unintended or unanticipated use or decisions relating to the way data are collected, represented, or used; lack of sensitivity to identity factors that contribute to bias in the evaluation of the algorithm, or misappropriation of the algorithm through miscommunicating or misunderstanding its limitations.
6	Diversity	The wide variety of shared and different personal and group characteristics among human beings. There are many kinds of diversity, including gender, sexual orientation, class, age, country of origin, education, religion, geography, physical or cognitive abilities, or other characteristics. Valuing diversity means recognizing differences between people, acknowledging that these differences are a valued asset, and striving for diverse representation as a critical step toward equity.
7	Inclusive	Avoiding bias by providing equitable and open access to opportunities and resources for engagement. This can be accomplished, for example, by enforcing fairness in the data collection methods, enforcing fairness in the assignment of labels, developing explainable, transparent, and interpretable models, having diverse teams monitor models, and looking for biases and eliminating them.
8	Fairness	Intent to promote nondiscrimination and population representation when assessing a group's eligibility for a benefit or penalty. This is particularly important given the statistical likelihood that artificial intelligence and machine learning systems could produce discriminatory outputs once algorithms are implemented across one or more data sets.
9	Representative	An individual or body chosen or appointed to act or speak for an individual, population, or subpopulation sharing a set of features or characteristics, including but not limited to gender, race, or sexual orientation.
10	Representative sample	A subset of a population that reflects the characteristics of the entire population from which it has been selected.
11	Gender identity	An individual's sense of oneself as male, female, or something else. When an individual's gender identity and biological sex are not congruent, the individual may identify along the transgender spectrum. An individual may choose to change their gender one or more times. Varying cultural indicators of gender, such as clothing choice, speech patterns, and personality traits, relate to gender but are not acceptable means to determine another's gender identity. The change in an individual's gender can be used to abuse, discriminate against, and misrepresent individuals and groups.
12	Sexual orientation	An individual's capacity for attraction to and sexual activity with the same or different sex. An individual's sexual orientation is indicated by one or more of the following: how an individual identifies their own sexual orientation, an individual's capacity for experiencing sexual and affectional attraction to people of the same or different gender, and an individual's sexual behavior with people of the same or different gender. Sexual orientation incorporates three core ideas: consensual human relationships—sexual, romantic, or both—the biological sex of an individual's actual or potential relationship partners, and enduring patterns of experience and behavior. Sexual minorities, or people whose sexual orientation does not conform to heteronormative cultural expectations, are vulnerable to violence and discrimination.

Figure 2. Definitional concepts of Artificial Intelligence/Machine Learning Consortium to Advance Health Equity and Researcher Diversity (AIM-AHEAD) ethics and equity glossary terms.



Interview Guide

As mentioned, extensive and iterative feedback received during the development of the quantitative survey led the EEWG coauthors and members to determine that a qualitative engagement approach is warranted to facilitate meaningful and diverse stakeholder engagement to disseminate and facilitate implementation of the principles and glossary. Therefore, the EEWG developed an interview guide that can be used or adapted to garner and understand AIM-AHEAD members' and other community perspectives on the principles and glossary. The interview guide is provided in [Multimedia Appendix 1](#).

Discussion

Overview

The role of those who will be affected by the findings of the research enterprise has evolved from their initial role as objects, as illustrated in the iconic painting of Edward Jenner administering the life-saving inoculation of the English boy with cowpox in 1796, the multi-episode television documentary "Microbes and Men," and the abuses of Black men in the US Public Health Service Study of the natural history of untreated syphilis at Tuskegee [30-33]. Over time, more attention has been devoted to assessing the potential harms and benefits of research to the people who are studied, albeit primarily as viewed by investigators, typically White men, and institutional review boards, typically comprised of researchers with minimal or latent community involvement. Incentivizing representation of nonscientific, nonaffiliate community members on institutional review boards, engaging members of historically underrepresented groups in more visible roles as investigators, and engaging minority-serving institutions as partners in AI and ML research is necessary to promote equitable access to opportunities and careers in AI and ML. Such an intentional

approach also, importantly, demonstrates an appreciation for local knowledge and facilitates the design of more culturally informed interventions that consider how research will affect heterogeneous populations being studied in AI and ML research. This form of appreciation is necessary for tailoring engagement to the needs of diverse groups and understanding how to overcome barriers to AI and ML research and use [34].

Beyond promoting diverse and equitable opportunities for participation in AI and ML research, it is necessary to recognize the need to translate that work into actual practice, which historically has also been a barrier to health equity. For example, the association of the lower-quality data measured by pulse oximetry with dark skin tones has long been known, and there have been versions of the technology designed to account for this discrepancy, but versions of pulse oximeters with biased tendencies remain in wide use [35]. There is a real risk that AI and ML technology will follow a similar pathway if there is not sufficient action to build ethics and equity into the research.

Overall, our effort reported here achieves 2 goals. The first is to describe what is needed procedurally and substantively to achieve equity. This is a complex process that must take place and evolve over time. It cannot be addressed as a 1-time event or by filling out a checklist. Achieving equity requires rebalancing the interests at stake in research, which, at a minimum, means truly considering and addressing the interests of the people who will be affected by the results. Ideally, research participants can become cocreators as ethics in AI and ML and related ethical principles evolve into more commonly accepted policies and practices. The second goal of this reported effort is to emphasize that addressing equity requires an inclusive, ongoing process with a shared understanding of salient terms that will evolve over time. Recent engagements within the AIM-AHEAD program have noted this to be true even for terms like AI and ML, as today very few stakeholders have been

able to clearly articulate how AI and ML can be or is used in the real world [34]. New and ongoing national initiatives, such as the National Academy of Medicine's AI Code of Conduct project, which intends to develop a "code of conduct for the development and use of AI in health, medical care, and health research," are encouraged to learn from the EEWG's efforts [36].

Our work builds on and can be incorporated into current AI and ML ethics and equity frameworks and policies within and outside of the United States, focused on improving population health through broad community involvement in AI and ML application development [17,36-38]. This includes, but is not limited to, the National Institutes of Health's policies and programs on AI and ML application development in health research; policy developments undertaken by the US Senate Health, Education, Labor, and Pensions Committee; the National Academy of Medicine's Artificial Intelligence Code of Conduct project; the European Commission's Guidelines for Trustworthy AI; Asilomar AI Principles; and lastly and importantly, a groundbreaking and recent US White House Executive Order explicitly supporting the mission of the AIM-AHEAD [36,39-42].

Importantly, our work provides a complementary, fundamental, and basic blueprint or process, along with operational tools and building blocks, to educate stakeholders on this practice of creating safe spaces and setting culture tones for diverse stakeholder engagement and consensus around best practices and shared terminology. Also importantly, our tools enable the collection of ongoing and iterative feedback concerning the local implementation of our principles and glossary. Iterations may be further disseminated, along with public-facing endorsements of the principles and glossary in their current form, by like-minded stakeholders seeking to ensure that researcher diversity, community, and social justice concerns influence AI and ML application development processes in health research and, broadly, science and technology.

Inclusive and ongoing processes to develop a shared understanding of salient terms like AI and ML and those described in our glossary require more time, greater inclusion, and deeper incorporation of diverse community perspectives. This approach differs drastically from the typical project life cycles afforded by the gold rush mentality that has emerged with AI and ML today. Therefore, one key step, moving forward, would be to persuade leaders in the AI and ML research enterprise to broadly disseminate the lessons that may be learned in operationalizing our EEWG principles and glossary. Programs such as the AIM-AHEAD need to objectively assess their administrative processes and evaluation criteria for what constitutes ethical and equitable opportunities for an AI and ML investigation, including investigator inclusion, data governance, data sources, and data infrastructure.

There are limitations to consider in our process and recommendations. First, the EEWG has continuously revisited the principles and glossary for potential editing based on the members' evolving experience and expert opinions, even though making these deliverables "living documents" complicates the

process of achieving sustainable consensus. Nonetheless, the principles and glossary will require reflection, appreciation, and adjustments over time to account for the effects of real-world events, human choices, or interpersonal phenomena from relevant perspectives. Also, some of our proposed glossary terms may already be limited in scope with respect to real-world events and phenomena. For instance, although our definition of "representative" concerns "an individual or body chosen or appointed to act or speak for an individual, population, or subpopulation," there are certain matters in which a representative may be self-appointed without specific authorization from those they wish to represent.

Therefore, ongoing engagement around the use of our principles and glossary in AI and ML research settings is encouraged to maximize their potential benefits and minimize any potential harm. However, ongoing engagement with institutions that have limited resources to support their full participation requires careful consideration and discussion of how to incentivize, support, and sustain meaningful engagement beyond mere compensation. One way to accomplish this is to seek institutional input through authentic connections to determine what they consider a valuable investment for their time, instead of deciding for them. For example, such connections can be made both within and outside of conferences, convenings, and events hosted by minority-serving institutions nationwide (eg, the Annual Biomedical Research Conference for Minoritized Scientists or the National Society of Black Engineers' Annual Convention).

Conclusions and Next Steps

An overemphasis on speed or velocity works against taking the time needed to foster the inclusion of historically and presently underrepresented communities in the development of AI and ML, ultimately rewarding AI and ML "haves" over "have-nots." In the private sector (eg, big technology companies and startups), the pace of AI and ML development is extremely rapid and difficult to manage. Inequitable divisions in access to resources like computers, smartphones, and the internet have vastly decreased over the past decade. Yet, AI and ML technology that is used with adequate operational know-how and e-literacy, cost of use, human resources and staffing needs to maintain cyberinfrastructure, and many other technical and nontechnical resources, is where these inequitable divisions can be addressed.

An equity-oriented public sector intervention, such as the AIM-AHEAD, can be more effective in achieving diversity and inclusion goals by emphasizing actions that do not sacrifice trust-building for the sake of rapid development of technology, especially in the initial stages. By slowing down to meet historically and presently underresourced institutions and communities where they are and where they are capable of engaging and competing, we can more effectively evaluate AI and ML implementation and results for bias over time and expand the potential to achieve the aims of ethics and equity. We envision a virtuous cycle of shared learning, building on our EEWG deliverables, that may bridge researchers and impacted communities into a new intersection of computational sciences, ethics, and health equity.

Acknowledgments

The authors would like to acknowledge Artificial Intelligence/Machine Learning Consortium to Advance Health Equity and Researcher Diversity (AIM-AHEAD) Ethics and Equity Workgroup (EEWG) members engaged during years 1 and 2 of the consortium.

The activities reported in this publication were supported by the Office of the Director, National Institutes of Health Common Fund, under award 1OT2OD032581. The funders had no role in study design, data collection and analysis, the decision to publish, or the preparation of the manuscript. More information can be found on the official website [43].

Conflicts of Interest

RHS is employed by the Duke-Margolis Center for Health Policy. BM is the Coeditor in Chief of JMIR AI but was excluded from the review of the manuscript. The other authors declare that they have no conflicts of interest.

Multimedia Appendix 1

Interview guide.

[DOCX File, 15 KB - ai_v2i1e52888_app1.docx]

References

1. Ayers JW, Poliak A, Dredze M, Leas EC, Zhu Z, Kelley JB, et al. Comparing physician and artificial intelligence chatbot responses to patient questions posted to a public social media forum. *JAMA Intern Med* 2023 Jun 01;183(6):589-596. [doi: [10.1001/jamainternmed.2023.1838](https://doi.org/10.1001/jamainternmed.2023.1838)] [Medline: [37115527](https://pubmed.ncbi.nlm.nih.gov/37115527/)]
2. Ross C. Google is training its generative AI to analyze medical images—and talk to doctors about them. *STAT*. 2023. URL: <https://www.statnews.com/2023/05/10/google-artificial-intelligence-ai-medpalm2-health/> [accessed 2023-07-30]
3. Vyas DA, Eisenstein LG, Jones DS. Hidden in plain sight—reconsidering the use of race correction in clinical algorithms. *N Engl J Med* 2020;383(9):874-882 [FREE Full text] [doi: [10.1056/NEJMms2004740](https://doi.org/10.1056/NEJMms2004740)] [Medline: [32853499](https://pubmed.ncbi.nlm.nih.gov/32853499/)]
4. Doshi RH, Bajaj S. Promises—and pitfalls—of ChatGPT-assisted medicine. *STAT*. 2023. URL: <https://www.statnews.com/2023/02/01/promises-pitfalls-chatgpt-assisted-medicine/> [accessed 2023-07-30]
5. Castillo A. Tools to predict stroke risk work less well for Black patients, study finds. *STAT*. 2023. URL: <https://www.statnews.com/2023/02/22/stroke-risk-machine-learning-models/> [accessed 2023-07-30]
6. Aggarwal R, Farag S, Martin G, Ashrafian H, Darzi A. Patient perceptions on data sharing and applying artificial intelligence to health care data: cross-sectional survey. *J Med Internet Res* 2021;23(8):e26162 [FREE Full text] [doi: [10.2196/26162](https://doi.org/10.2196/26162)] [Medline: [34236994](https://pubmed.ncbi.nlm.nih.gov/34236994/)]
7. Timmons AC, Duong JB, Simo Fiallo N, Lee T, Vo HPQ, Ahle MW, et al. A call to action on assessing and mitigating bias in artificial intelligence applications for mental health. *Perspect Psychol Sci* 2023;18(5):1062-1096 [FREE Full text] [doi: [10.1177/17456916221134490](https://doi.org/10.1177/17456916221134490)] [Medline: [36490369](https://pubmed.ncbi.nlm.nih.gov/36490369/)]
8. Lee M, Rich K. Who is included in human perceptions of AI?: trust and perceived fairness around healthcare AI and cultural mistrust. New York, NY, US: Association for Computing Machinery; 2021 Presented at: CHI '21: Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems; May 8-13, 2021; Yokohama, Japan p. 1-14. [doi: [10.1145/3411764.3445570](https://doi.org/10.1145/3411764.3445570)]
9. Smith SS. Race and trust. *Annu Rev Sociol* 2010;36(1):453-475. [doi: [10.1146/annurev.soc.012809.102526](https://doi.org/10.1146/annurev.soc.012809.102526)]
10. Hai AA, Weiner MG, Paranjape A, Livshits A, Brown JR, Obradovic Z, et al. Deep learning vs traditional models for predicting hospital readmission among patients with diabetes. *AMIA Annu Symp Proc* 2022;2022:512-521 [FREE Full text] [Medline: [37128461](https://pubmed.ncbi.nlm.nih.gov/37128461/)]
11. Rubin DJ, Gogineni P, Deak A, Vaz C, Watts S, Recco D, et al. The Diabetes Transition of Hospital Care (DiaTOHC) pilot study: a randomized controlled trial of an intervention designed to reduce readmission risk of adults with diabetes. *J Clin Med* 2022;11(6):1471 [FREE Full text] [doi: [10.3390/jcm11061471](https://doi.org/10.3390/jcm11061471)] [Medline: [35329797](https://pubmed.ncbi.nlm.nih.gov/35329797/)]
12. Cerrato P, Halamka J, Pencina M. A proposal for developing a platform that evaluates algorithmic equity and accuracy. *BMJ Health Care Inform* 2022;29(1):e100423 [FREE Full text] [doi: [10.1136/bmjhci-2021-100423](https://doi.org/10.1136/bmjhci-2021-100423)] [Medline: [35410952](https://pubmed.ncbi.nlm.nih.gov/35410952/)]
13. Rising Equitable Community Data Ecosystems (RECoDE). data.org. 2022. URL: <https://data.org/reports/recode-report/> [accessed 2023-07-30]
14. The proliferation of AI ethics principles: what's next? Montreal AI Ethics Institute. 2021. URL: <https://montrealthics.ai/the-proliferation-of-ai-ethics-principles-whats-next/> [accessed 2023-07-30]
15. Rajkomar A, Hardt M, Howell MD, Corrado G, Chin MH. Ensuring fairness in machine learning to advance health equity. *Ann Intern Med* 2018;169(12):866-872 [FREE Full text] [doi: [10.7326/M18-1990](https://doi.org/10.7326/M18-1990)] [Medline: [30508424](https://pubmed.ncbi.nlm.nih.gov/30508424/)]
16. Wiens J, Saria S, Sendak M, Ghassemi M, Liu VX, Doshi-Velez F, et al. Do no harm: a roadmap for responsible machine learning for health care. *Nat Med* 2019;25(9):1337-1340. [doi: [10.1038/s41591-019-0548-6](https://doi.org/10.1038/s41591-019-0548-6)] [Medline: [31427808](https://pubmed.ncbi.nlm.nih.gov/31427808/)]

17. Dankwa-Mullan I, Scheufele EL, Matheny ME, Quintana Y, Chapman WW, Jackson G, et al. A proposed framework on integrating health equity and racial justice into the artificial intelligence development lifecycle. *J Health Care Poor Underserved* 2021;32(2):300-317. [doi: [10.1353/hpu.2021.0065](https://doi.org/10.1353/hpu.2021.0065)]
18. Vayena E, Blasimme A, Cohen IG. Machine learning in medicine: addressing ethical challenges. *PLoS Med* 2018;15(11):e1002689 [FREE Full text] [doi: [10.1371/journal.pmed.1002689](https://doi.org/10.1371/journal.pmed.1002689)] [Medline: [30399149](https://pubmed.ncbi.nlm.nih.gov/30399149/)]
19. Chi N, Lurie E, Mulligan D. Reconfiguring diversity and inclusion for AI ethics. 2021 Presented at: Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society; May 19-21, 2021; USA p. 447-457. [doi: [10.1145/3461702.3462622](https://doi.org/10.1145/3461702.3462622)]
20. Karnik NS, Afshar M, Churpek MM, Nunez-Smith M. Structural disparities in data science: a prolegomenon for the future of machine learning. *Am J Bioeth* 2020;20(11):35-37 [FREE Full text] [doi: [10.1080/15265161.2020.1820102](https://doi.org/10.1080/15265161.2020.1820102)] [Medline: [33103976](https://pubmed.ncbi.nlm.nih.gov/33103976/)]
21. UNESCO member states adopt the first ever global agreement on the ethics of artificial intelligence. UNESCO. URL: <https://www.unesco.org/en/articles/unesco-member-states-adopt-first-ever-global-agreement-ethics-artificial-intelligence> [accessed 2023-07-30]
22. McLennan S, Lee MM, Fiske A, Celi LA. AI ethics is not a panacea. *Am J Bioeth* 2020;20(11):20-22 [FREE Full text] [doi: [10.1080/15265161.2020.1819470](https://doi.org/10.1080/15265161.2020.1819470)] [Medline: [33103983](https://pubmed.ncbi.nlm.nih.gov/33103983/)]
23. Jobin A, Ienca M, Vayena E. The global landscape of AI ethics guidelines. *Nat Mach Intell* 2019;1(9):389-399. [doi: [10.1038/s42256-019-0088-2](https://doi.org/10.1038/s42256-019-0088-2)]
24. Future of privacy forum resources—ethics, governance, and compliance resources. URL: <https://sites.google.com/fpf.org/futureofprivacyforumresources/ethics-governance-and-compliance-resources> [accessed 2023-07-30]
25. Morley J, Machado CCV, Burr C, Cowls J, Joshi I, Taddeo M, et al. The ethics of AI in health care: a mapping review. *Soc Sci Med* 2020;260:113172. [doi: [10.1016/j.socscimed.2020.113172](https://doi.org/10.1016/j.socscimed.2020.113172)] [Medline: [32702587](https://pubmed.ncbi.nlm.nih.gov/32702587/)]
26. Wilkins C. Effective engagement requires trust and being trustworthy. *Med Care* 2018;56(10 Suppl 1):S6-S8 [FREE Full text] [doi: [10.1097/MLR.0000000000000953](https://doi.org/10.1097/MLR.0000000000000953)] [Medline: [30015725](https://pubmed.ncbi.nlm.nih.gov/30015725/)]
27. Glover WJ, Hendricks-Sturup R. Ethics and equity-centred perspectives in engineering systems design. In: Maier A, Oehmen J, Vermaas PE, editors. *Handbook of Engineering Systems Design*. London: Springer Cham; 2022:1-24.
28. Dalkey N. An experimental study of group opinion. *Futures* 1969;1(5):408-426. [doi: [10.1016/s0016-3287\(69\)80025-x](https://doi.org/10.1016/s0016-3287(69)80025-x)]
29. Nasa P, Jain R, Juneja D. Delphi methodology in healthcare research: how to decide its appropriateness. *World J Methodol* 2021;11(4):116-129 [FREE Full text] [doi: [10.5662/wjm.v11.i4.116](https://doi.org/10.5662/wjm.v11.i4.116)] [Medline: [34322364](https://pubmed.ncbi.nlm.nih.gov/34322364/)]
30. Riedel S. Edward Jenner and the history of smallpox and vaccination. *Proc (Bayl Univ Med Cent)* 2005;18(1):21-25 [FREE Full text] [doi: [10.1080/08998280.2005.11928028](https://doi.org/10.1080/08998280.2005.11928028)] [Medline: [16200144](https://pubmed.ncbi.nlm.nih.gov/16200144/)]
31. Strickler DA. Microbes and men. *JAMA* 1916;LXVI(1):52. [doi: [10.1001/jama.1916.02580270056031](https://doi.org/10.1001/jama.1916.02580270056031)]
32. Brandt AM. Racism and research: the case of the Tuskegee Syphilis study. *Hastings Cent Rep* 1978;8(6):21-29. [doi: [10.2307/3561468](https://doi.org/10.2307/3561468)]
33. The untreated syphilis study at Tuskegee timeline. Centers for Disease Control and Prevention. 2022. URL: <https://www.cdc.gov/tuskegee/timeline.htm> [accessed 2023-08-27]
34. Vishwanatha JK, Christian A, Sambamoorthi U, Thompson EL, Stinson K, Syed TA. Community perspectives on AI/ML and health equity: AIM-AHEAD nationwide stakeholder listening sessions. *PLOS Digit Health* 2023;2(6):e0000288 [FREE Full text] [doi: [10.1371/journal.pdig.0000288](https://doi.org/10.1371/journal.pdig.0000288)] [Medline: [37390116](https://pubmed.ncbi.nlm.nih.gov/37390116/)]
35. Zou J, Schiebinger L. Ensuring that biomedical AI benefits diverse populations. *EBioMedicine* 2021;67:103358 [FREE Full text] [doi: [10.1016/j.ebiom.2021.103358](https://doi.org/10.1016/j.ebiom.2021.103358)] [Medline: [33962897](https://pubmed.ncbi.nlm.nih.gov/33962897/)]
36. Health care artificial intelligence code of conduct. National Academy of Medicine. URL: <https://nam.edu/programs/value-science-driven-health-care/health-care-artificial-intelligence-code-of-conduct/> [accessed 2023-07-31]
37. Advancing health care AI through ethics, evidence and equity. American Medical Association. 2023. URL: <https://www.ama-assn.org/practice-management/digital/advancing-health-care-ai-through-ethics-evidence-and-equity> [accessed 2023-10-25]
38. Berdahl CT, Baker L, Mann S, Osoba O, Giroso F. Strategies to improve the impact of artificial intelligence on health equity: scoping review. *JMIR AI* 2023;2(1):e42936 [FREE Full text]
39. Ranking member cassidy releases white paper on artificial intelligence. The U.S. Senate Committee on Health, Education, Labor & Pensions. 2023. URL: <https://www.help.senate.gov/ranking/newsroom/press/ranking-member-cassidy-releases-white-paper-on-artificial-intelligence> [accessed 2023-10-25]
40. Ethics guidelines for trustworthy AI. European Commission: Shaping Europe's digital future. 2019. URL: <https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai> [accessed 2023-10-25]
41. AI principles. Future of Life Institute. 2017. URL: <https://futureoflife.org/open-letter/ai-principles/> [accessed 2023-10-25]
42. Executive order on the safe, secure, and trustworthy development and use of artificial intelligence. The White House. 2023. URL: <https://www.whitehouse.gov/briefing-room/presidential-actions/2023/10/30/executive-order-on-the-safe-secure-and-trustworthy-development-and-use-of-artificial-intelligence/> [accessed 2023-10-31]
43. AIM-AHEAD. National Institutes of Health: Office of Data Science Strategy. URL: <https://datascience.nih.gov/artificial-intelligence/aim-ahead> [accessed 2023-11-21]

Abbreviations

AI: artificial intelligence

AIM-AHEAD: Artificial Intelligence/Machine Learning Consortium to Advance Health Equity and Researcher Diversity

EEWG: Ethics and Equity Workgroup

ML: machine learning

Edited by A Mavragani; submitted 18.09.23; peer-reviewed by S Wiertz, S Bito; comments to author 16.10.23; revised version received 01.11.23; accepted 05.11.23; published 06.12.23.

Please cite as:

*Hendricks-Sturup R, Simmons M, Anders S, Aneni K, Wright Clayton E, Coco J, Collins B, Heitman E, Hussain S, Joshi K, Lemieux J, Lovett Novak L, Rubin DJ, Shanker A, Washington T, Waters G, Webb Harris J, Yin R, Wagner T, Yin Z, Malin B
Developing Ethics and Equity Principles, Terms, and Engagement Tools to Advance Health Equity and Researcher Diversity in AI and Machine Learning: Modified Delphi Approach*

JMIR AI 2023;2:e52888

URL: <https://ai.jmir.org/2023/1/e52888>

doi: [10.2196/52888](https://doi.org/10.2196/52888)

PMID: [38875540](https://pubmed.ncbi.nlm.nih.gov/38875540/)

©Rachele Hendricks-Sturup, Malaika Simmons, Shilo Anders, Kammarauche Aneni, Ellen Wright Clayton, Joseph Coco, Benjamin Collins, Elizabeth Heitman, Sajid Hussain, Karuna Joshi, Josh Lemieux, Laurie Lovett Novak, Daniel J Rubin, Anil Shanker, Talitha Washington, Gabriella Waters, Joyce Webb Harris, Rui Yin, Teresa Wagner, Zhijun Yin, Bradley Malin. Originally published in JMIR AI (<https://ai.jmir.org>), 06.12.2023. This is an open-access article distributed under the terms of the Creative Commons Attribution License (<https://creativecommons.org/licenses/by/4.0/>), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR AI, is properly cited. The complete bibliographic information, a link to the original publication on <https://www.ai.jmir.org/>, as well as this copyright and license information must be included.

Original Paper

A Scalable Radiomics- and Natural Language Processing–Based Machine Learning Pipeline to Distinguish Between Painful and Painless Thoracic Spinal Bone Metastases: Retrospective Algorithm Development and Validation Study

Hossein Naseri¹, MSc; Sonia Skamene², MD; Marwan Tolba², MD; Mame Daro Faye², MD; Paul Ramia², MD; Julia Khriouian², MD; Marc David², MD; John Kildea¹, PhD

¹Medical Physics Unit, McGill University Health Centre, Montreal, QC, Canada

²Division of Radiation Oncology, McGill University Health Centre, Montreal, QC, Canada

Corresponding Author:

Hossein Naseri, MSc
Medical Physics Unit
McGill University Health Centre
Cedars Cancer Centre
1001 boul Décarie Montréal
Montreal, QC, H4A 3J1
Canada
Phone: 1 514 934 1934 ext 44158
Email: 3naseri@gmail.com

Abstract

Background: The identification of objective pain biomarkers can contribute to an improved understanding of pain, as well as its prognosis and better management. Hence, it has the potential to improve the quality of life of patients with cancer. Artificial intelligence can aid in the extraction of objective pain biomarkers for patients with cancer with bone metastases (BMs).

Objective: This study aimed to develop and evaluate a scalable natural language processing (NLP)– and radiomics-based machine learning pipeline to differentiate between painless and painful BM lesions in simulation computed tomography (CT) images using imaging features (biomarkers) extracted from lesion center–based regions of interest (ROIs).

Methods: Patients treated at our comprehensive cancer center who received palliative radiotherapy for thoracic spine BM between January 2016 and September 2019 were included in this retrospective study. Physician-reported pain scores were extracted automatically from radiation oncology consultation notes using an NLP pipeline. BM center points were manually pinpointed on CT images by radiation oncologists. Nested ROIs with various diameters were automatically delineated around these expert-identified BM center points, and radiomics features were extracted from each ROI. Synthetic Minority Oversampling Technique resampling, the Least Absolute Shrinkage And Selection Operator feature selection method, and various machine learning classifiers were evaluated using precision, recall, F_1 -score, and area under the receiver operating characteristic curve.

Results: Radiation therapy consultation notes and simulation CT images of 176 patients (mean age 66, SD 14 years; 95 males) with thoracic spine BM were included in this study. After BM center point identification, 107 radiomics features were extracted from each spherical ROI using pyradiomics. Data were divided into 70% and 30% training and hold-out test sets, respectively. In the test set, the accuracy, sensitivity, specificity, and area under the receiver operating characteristic curve of our best performing model (neural network classifier on an ensemble ROI) were 0.82 (132/163), 0.59 (16/27), 0.85 (116/136), and 0.83, respectively.

Conclusions: Our NLP- and radiomics-based machine learning pipeline was successful in differentiating between painful and painless BM lesions. It is intrinsically scalable by using NLP to extract pain scores from clinical notes and by requiring only center points to identify BM lesions in CT images.

(JMIR AI 2023;2:e44779) doi:[10.2196/44779](https://doi.org/10.2196/44779)

KEYWORDS

cancer; pain; palliative care; radiotherapy; bone metastases; radiomics; natural language processing; machine learning; artificial intelligent; radiation therapy

Introduction

Overview

Most patients with cancer with bone metastasis (BM) experience pain [1] and most receive radiotherapy to control it [2]. But, it has been shown that due to the subjective and qualitative nature of the pain, clinicians often underestimate pain [3]. As a result, many patients with BM receive radiotherapy after their pain has already become debilitating [4].

Although patient-reported outcomes can be used to obtain pain scores directly from patients themselves, the efficacy of these pain scores is limited due to the fact that these ratings are highly qualitative and subjective [5]. Because of this, it is desirable to have pain scoring systems that are more objective. The goal of this study was to explore ways to automatically and objectively quantify pain associated with BMs using computed tomography (CT) images.

We hypothesized that tumor features extracted from CT images of BMs contain imaging biomarkers that may be used to objectively identify BM-associated pain. These pain biomarkers may provide the opportunity to develop objective pain scoring tools to aid in the diagnosis, treatment, understanding, and prognosis of BM pain.

Background

The search for imaging and nonimaging pain biomarkers has been the focus of numerous studies [5-12]. Various studies [13-21] have shown how artificial intelligence (AI), including machine learning and radiomics, can be used to understand and quantify pain. For example, Mashayekhi et al [22] showed that radiomic features extracted from the CT images of the pancreas can help to identify functional abdominal pain in patients. Vedantam et al [23] explored the viability of using radiomics features extracted from magnetic resonance images to detect pain following percutaneous cordotomy. At least 1 study [13] has reported using radiomics to identify painful metastatic lesions in radiographic images. However, we found no reports in the literature of a scalable approach that can be used efficiently on a large set of unlabeled patient data. To the best of our knowledge, our work is the first to combine natural language processing (NLP) and radiomics to enable an efficient and scalable pain identification pipeline using unstructured data.

A fundamental challenge in developing any AI model for use in medicine is the need to obtain sufficient patient data for training and testing. For example, the data set used by Wakabayashi et al in the study that we mentioned earlier [13], was limited to 69 patients. One limiting factor is obtaining standard patient-reported pain scores for use as ground-truth data, and another limiting factor is obtaining segmented images from which to extract tumor biomarkers. For the work reported in this paper, we overcame the data set size limitation by using 2 novel strategies. First, by combining NLP with radiomics, we quickly mined pain scores from clinical notes and used these

NLP-extracted scores to label our radiomics features for supervised learning. Second, by asking our clinical colleagues to pinpoint only the center points of BM lesions in radiotherapy simulation CT images, we maximized the number of lesions identified in the time available. In the medical field, NLP has shown promising results in extracting biomedical information and clinical outcomes such as pain from unstructured text data [24-26]. Moreover, as we reported previously [21], by automatically delineating geometrical regions around BM lesion center points, it is possible to successfully extract radiomics features for robust BM lesion detection. In this study, we report how our combined radiomics-NLP machine learning pipeline can successfully identify pain in radiotherapy simulation CT images of patients with cancer with BMs.

Methods

Ethical Considerations

This retrospective study was approved by the research ethics board of the McGill University Health Centre (2020-5899) with the waiver of informed consent. We confirm that the entire research was performed in accordance with research ethics board's guidelines and regulations.

Data Selection

Our patient-selection process is outlined in [Figure 1](#). The initial number of 200 pairs of radiation oncology consultation notes and CT images of patients with spinal BM were included in this study based on the minimum sample size calculation as explained in Section A.1 in [Multimedia Appendix 1](#) [27]. In total, 120 of the notes and all 200 of the CT images from this study were independently used in 2 studies we previously reported on [21-25]. The first [25] of these studies showed the feasibility of extracting pain from consultation notes of patients with cancer, using NLP. The second [21] demonstrated the feasibility of using lesion center point-based radiomics models to differentiate healthy and metastatic bone lesions in CT scans of patients with BMs. This study combined the data and results from these 2 prior studies and expanded upon them to build an NLP- and radiomics-based model to detect pain using the CT scans of patients.

We searched our institution's Oncology Information System for the radiotherapy plans of patients diagnosed with a "secondary malignant neoplasm of bone" between January 2016 and September 2019. From the retrieved list, we selected those who were treated for thoracic spinal BM. Then, we retrieved the corresponding consultation notes and simulation CT images. A note-image pair was included if (1) the note was in English, (2) pain was documented, (3) the simulation CT image was taken up to 10 days post consultation, and (4) simulation CT revealed BM lesions in the thoracic spine. Patients with multiple but nonoverlapping note-image pairs were considered independent samples. We only considered the same patients as new participants if they had CT scans and associated

consultation notes for BM lesions in different areas of their spines. As a result, each BM lesion was included only once in our study. Also, it should be noted that palliative patients normally have their simulation CT scan (for treatment planning) on the same day or within a few days after the consultation, and radiotherapy is delivered on the same day or within a few days after treatment planning. To assure that there is no change in the BM lesion structure or pain status, we did not allow more than a 10-day gap between the two. Figure A1 in [Multimedia](#)

[Appendix 1](#) displays the distribution of the time interval between the radiotherapy consultation and CT acquisition dates.

We randomly assigned note-image pairs to the training or cross-validation set (approximately 70%) or the holdout test set (approximately 30%). We used stratified randomization to preserve the original sample ratio between pain labels in each sample set. In addition, we performed a paired *t* test and a chi-square analysis [28] to ensure that there was no systematic bias in any of our sample sets regarding gender, age, or primary cancer type. Patient demographics are presented in [Table 1](#).

Figure 1. The patient selection criteria used to obtain the radiotherapy consultation notes and simulation computed tomography (CT) images that formed our training and test data sets. The initial number of 200 note-image pairs included in this study was based on the minimum sample size calculation as explained in Section A.1 in [Multimedia Appendix 1](#). BM: bone metastases; DICOM: Digital Imaging and Communications in Medicine; RT: radiotherapy; T-spine: thoracic spine. *Four patients had pairs in both the training and test sets.

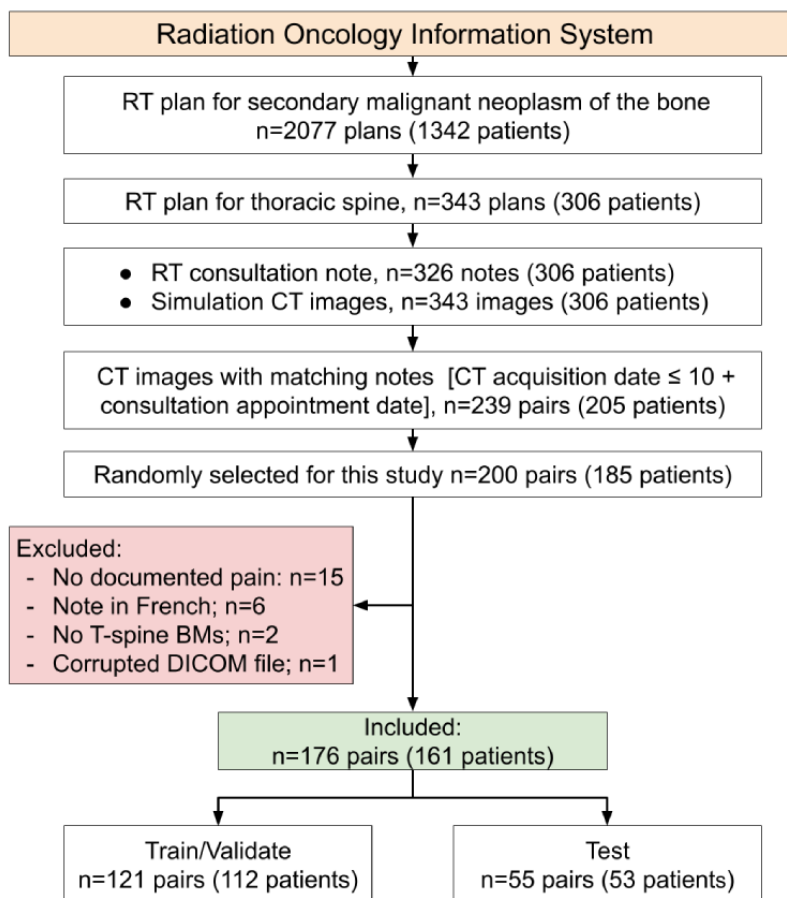


Table 1. Patient demographics in the training and test sets.

Characteristics	Training and validation set (n=121)	Test set (n=55)	<i>P</i> value ^a
Gender, n (%)			N/A ^b
Female	56 (46)	25 (45)	
Male	65 (54)	30 (55)	
Age (years), mean (SD)^c			N/A
Female	63 (14)	64 (12)	.99
Male	67 (14)	64 (13)	.72
Primary cancer type, n (%)			.06
Lung	32 (26)	20 (36)	
Breast	23 (19)	11 (20)	
Prostate	19 (16)	5 (9)	
Multiple myeloma	8 (7)	6 (11)	
Renal cell carcinoma	7 (6)	2 (4)	
Other and unknown	64 (53)	31 (56)	
Bone metastasis lesions, n (%)			.42
Lytic	220 (52)	76 (47)	
Blastic	122 (29)	57 (35)	
Mix	81 (19)	30 (18)	
Pain label, n (%)			N/A
Pain	357 (84)	136 (83)	
No pain	66 (16)	27 (17)	

^a*P* values for numerical values (age) and categorical features (primary cancer site and bone metastasis lesion type) were calculated using a 2-tailed heteroscedastic *t* test and a chi-square test, respectively.

^bN/A: not applicable.

^cThe *P* value for the age difference between males and females was .20 for the training and validation set and .50 for the test set.

NLP-Extracted Pain Labels

Due to the absence of patient-reported pain scores in our Oncology Information System, we extracted physician-reported pain scores from patients' radiation oncology consultation notes using our previously reported NLP pipeline [25]. While pain scores were typically reported as part of the "history of the present illness" in our hospital, for the sake of generalizability, we extracted pain scores from the entire note.

Our NLP pipeline first processed the text with MetaMap [29] and mapped it to the UMLS (ie, Unified Medical Language System) Metathesaurus [30] in order to identify pain terminologies and their severity scores. Next, it applied rules to filter out hypothetical, conditional, and historical references to pain in order to focus solely on references to pain at the time of the consultation. Then, it calculated the average pain intensity (API) in each note by averaging the pain scores therein. Finally, it assigned each note a "verbally declared pain" (VDP) label, as VDP="no pain" (if API0), and VDP="pain" (if API0). These pain labels were used to train, validate, and test our radiomics model.

Expert-Extracted Pain Scores

To evaluate the effect of NLP-extracted pain labels on the performance of our pipeline, we also generated best-available ground-truth pain labels using expert-annotated pain scores. To do so, our radiation oncologists used the texTRACTOR [31] pain labeling application to manually read consultation notes and label valid pain scores in our training and test data sets using a 4-grade verbal rating scale (no pain, mild, moderate, and severe). A mention of pain was regarded as valid if it reflected the status of pain at the metastatic sites for which treatment was planned at the time of the consultation. Table A1 in [Multimedia Appendix 1](#) contains all the NLP- and expert-extracted pain scores, and Figure A2 in [Multimedia Appendix 1](#) illustrates the level of agreement between them. Due to the quality of the documented pain scores and lack of interrater agreement among experts (Fleiss $\kappa=0.43$), as explained by Naseri et al [25], we subsequently defined a binary pain score as "no pain" and "pain" in order to establish satisfactory interrater agreement ($\kappa=0.66$) [25]. To create binary ground-truth pain labels comparable to the NLP-extracted labels, we assigned notes scored as "no pain" to "no pain" and notes scored as "mild," "moderate," and "severe" pain to "pain." These

expert-extracted pain scores were used to measure how well the NLP pipeline works.

Center Point Identification of BM Lesions

BM lesion center points were identified by a team comprising a staff radiation oncologist (SS) with 10 years' experience, a radiation oncology fellow (MT), and 3 third-year radiation oncology residents (J Khriguian, PR, and MF). Simulation CT DICOM (ie, Digital Imaging and Communications in Medicine) files were exported from the radiotherapy treatment planning software and deidentified. Then, the CT images were randomly divided into 5 sets and loaded into the diCOMBINE [32] application for BM lesion center point identification. Our experts were blinded to patients' pain statuses and identities. We requested each expert to label center points for all visually identifiable BM lesions in all CT images within 1 of the 5 sets, and another expert was assigned to validate their labels. A key benefit of this radiomics pipeline [21] is that it does not require full lesion segmentation, making it feasible to engage busy clinicians.

Segmentation of Regions of Interest

Using our previously reported methodology [21], we automatically segmented lesion center point-based nested

spherical (SP) regions of interest (ROIs). To do this, we first delineated nested spherical ROIs around the identified BM lesion center points (see [Textbox 1](#), top panel). ROI diameters ranged from 7 mm (3×3 voxels) to 50 mm (average size of the vertebral body) [33]. Then, in addition to what was reported by Naseri et al [21], we used Hounsfield units thresholding to exclude fat and air regions from the delineated ROIs. For this, motivated by Deglint et al [34] and Ulano et al [35], we applied a threshold to remove voxels with negative Hounsfield units from our ROIs. Hounsfield units of <0 are associated with fat and air [34]. We used OpenCV [36] (version 4.4.0) for Hounsfield units thresholding and applied a Gaussian filter to reduce noise. Then, we used pynrrd [37] (version 0.4.2) to export each ROI as a 3D binary mask and store it as a.nrrd [38] file. Finally, we aggregated these nested ROI masks to form ensemble ROIs. In this study, we examined 2 contrasting ensemble (EN) ROIs as shown in [Textbox 1](#) (bottom panel): one with small size and 3 layers (EN3) and the other with large size and 6 layers (EN6). Wakabayashi et al [13] and Naseri et al [21] have shown that radiomics-based machine learning models trained on ensemble ROIs have better classification performance than single ROI-based models.

Textbox 1. The characteristics of the spherical and ensemble regions of interest (ROIs) used in this study.

Nested spherical (SP) ROIs with Hounsfield units (HUs) intensity thresholds (HU>0):

- SP7 (diameter 7 mm)
- SP10 (diameter 10 mm)
- SP15 (diameter 15 mm)
- SP20 (diameter 20 mm)
- SP30 (diameter 30 mm)
- SP50 (diameter 50 mm)

Ensemble (EN) ROIs:

- EN3 (ROI SP7+SP10+SP15)
- EN6 (ROI SP7+SP10+SP15+SP20+SP30+SP50)

Radiomics Models

Our radiomics pipeline is illustrated in [Figure 2](#). We essentially used our previously reported pipeline [21] but with our NLP- and expert-extracted pain labels to train and test it. We made one improvement to the pipeline by incorporating Imbalanced-learn [39] (version 0.7.0) as a resampling step to account for imbalance (see below).

Radiomics features were extracted from each CT image using masks composed of the ensemble ROIs listed in [Textbox 1](#). Then, the feature space was scaled using *z* score normalization [40], and the associated NLP-extracted binary pain labels (pain=1, no pain=0) were incorporated. A single NLP-extracted pain score was assigned to all the lesions extracted from a given paired CT image.

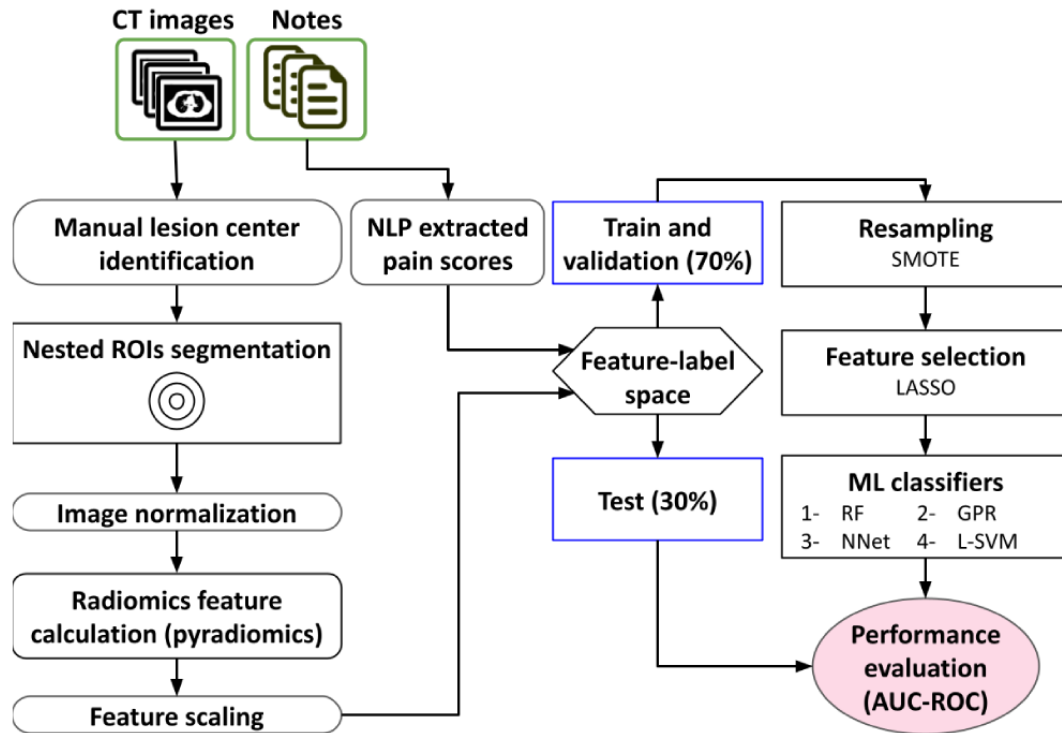
Due to the nature of BM pain [41], there was a large imbalance between the number of painful and painless lesions (493 pain,

93 no pain). Therefore, we used the Synthetic Minority Oversampling Technique (SMOTE) [42] in the training phase as it has been shown to be the best-performing resampling method for radiomics [43]. We did not apply resampling to our test set in order to maintain the original sample imbalance. Then, the Least Absolute Shrinkage And Selection Operator [44] feature selection method was applied to the feature space to remove noninformative features. Least Absolute Shrinkage And Selection Operator is a commonly used feature selection method in radiomics studies [45,46]. Finally, we examined the Gaussian process regression, linear support vector machine, random forest, and neural networks classifiers, as they were the best performing machine learning classifiers in our previous work. We evaluated the performance of our models on the training set using 5-fold cross-validation. Final evaluation was performed on the test set. The receiver operating characteristic (ROC) [47] curve, area under the ROC curve (AUC), precision, sensitivity, specificity, and F_1 -score metrics were used to report the performance of

our models on the training and test sets. We also trained and tested our best performing pipeline using the expert-extracted

pain scores (best-available ground-truth) to evaluate the impact of NLP-extracted pain labels.

Figure 2. The radiomics-based pipeline that we used to select and train a machine learning model to separate painful and painless bone metastasis lesions. Our pipeline is the same as that published by Naseri et al [21] but using NLP-extracted pain labels and modified to account for sample imbalance. AUC-ROC: area under the receiver operating characteristic curve-receiver operating characteristic; CT: computed tomography; GPR: Gaussian process regression; LASSO: Least Absolute Shrinkage And Selection Operator; L-SVM: linear support vector machine; ML: machine learning; NLP: natural language processing; NNet: neural network; RF: random forest; ROI: region of interest; SMOTE: Synthetic Minority Oversampling Technique.



Results

Patient Demographics

A total of 176 pairs of radiotherapy consultation notes and simulation CT images of patients with thoracic spinal BM were included in this study. As summarized in Table 1, a total of 121 sample pairs (mean patient age 63, SD 14 years; males: n=65, mean age 67, SD 14 years; $P=.20$) were included for training and cross-validation, and 55 sample pairs (mean patient age 64, SD 12 years; males: n=25, mean age 64, SD 13 years; females: mean age 64, SD 23 years; $P=.50$) were included in the test set. The sample selection procedure and data quantities are presented in Figure 1. The demographics of the patients in the training and test sets are presented in Table 1. The most common primary

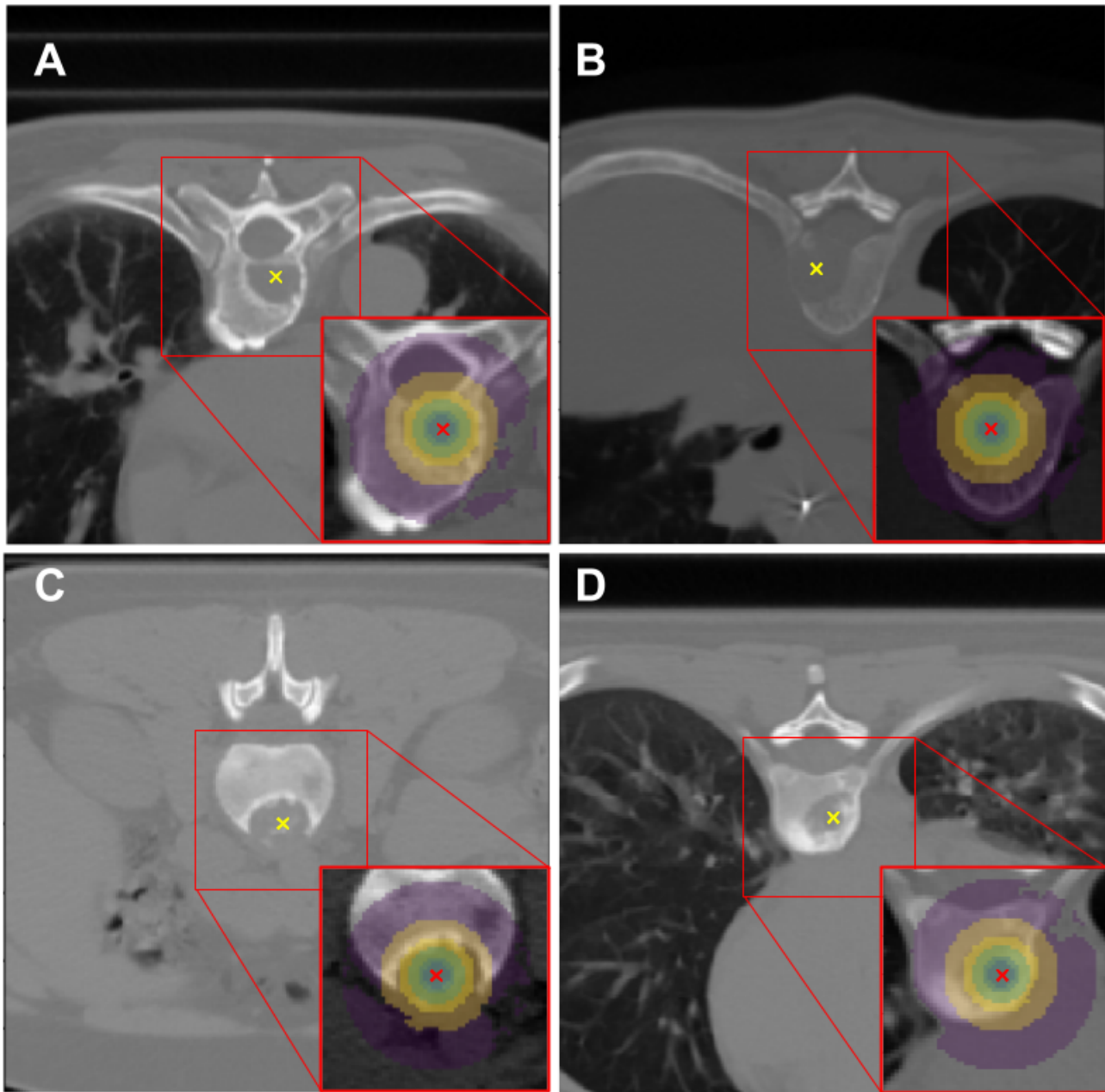
cancer sites were the lungs (n=52), breasts (n=34), and prostate (n=24).

A total of 586 BM center points were identified by our experts on the training (n=423 lesions) and test (n=163 lesions) data sets. In the training set, 357 (84%) lesions were labeled by the NLP pipeline as painful and 66 lesions were labeled as painless. In the test set, 136 (83%) lesions were identified by the NLP pipeline as painful, and 27 lesions were labeled as painless. This represented a significant but equal imbalance in our training and test sets.

Segmented ROIs

Examples of segmented ROIs with the Hounsfield units threshold applied are presented in Figure 3 for painful and painless BMs.

Figure 3. Examples of segmented nested spherical regions of interest (ROIs) with the Hounsfield units threshold applied on computed tomography images of patients with painful (A, B) and painless (C, D) bone metastases lesions. Nested ROIs with diameters of 50, 30, 20, 15, 10, and 7 mm are shown in the insets as different hues.



Testing Our Radiomics Models

In total, 107 radiomics features were extracted from each of the 6 nested ROIs. Then, they were aggregated to form feature spaces for the EN3 (with 321 features) and EN6 (with 642 features) ensemble ROIs. Figure 4 shows the ROC curve of each model in the training (black lines) and test (red squares) data sets using the EN3 and EN6 ROIs. On the training set, the gray range represents the mean (SD) AUC of the 5-fold

cross-validation. The AUC and F_1 -score grids are presented in Table 2.

The precision, accuracy, sensitivity, specificity, F_1 -score, and AUC values of our best-performing pipeline (neural networks with the EN6 ROI) are presented in Table 3. The performance of this pipeline (trained and tested) on the data set of expert-extracted pain labels (best-available ground-truth) is provided as a quality measurement. The performance of the model described previously by Wakabayashi et al [13] is also provided for comparison.

Figure 4. Receiver operating characteristic curves for our classifiers using 3-layer ensemble (EN3) (top row) and 6-layer ensemble (EN6) (bottom row) lesion center point–based ensemble regions of interest in training (black lines) and test (dark red squares) data sets. AUC: area under the receiver operating characteristic curve; GPR: Gaussian process regression; L-SVM: linear support vector machine; NNet: neural network; RF: random forest.

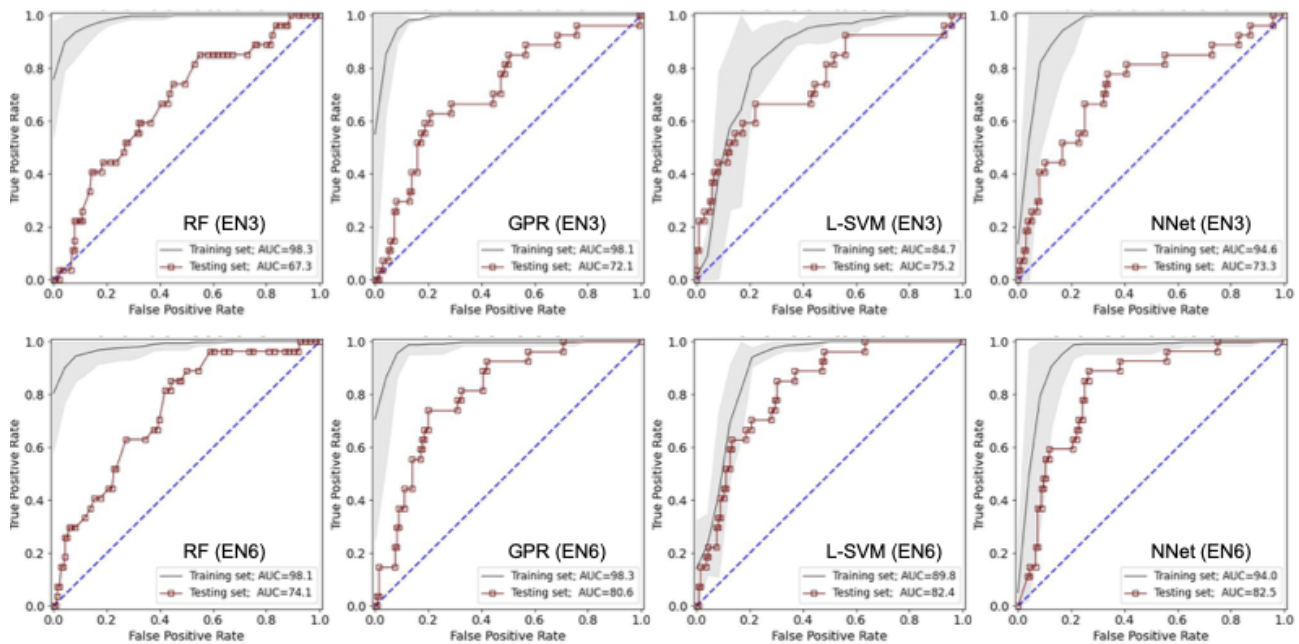


Table 2. The area under the receiver operating characteristic curves (AUCs) and F_1 -scores of our machine learning classifiers in the training and test data sets using the ensemble (EN) regions of interest EN3 and EN6 for each of the RF (random forest), GPR (Gaussian process regression), L-SVM (linear support vector machine), and NNet (neural networks) classifiers.

Region of interest	Training set				Test set			
	RF	GPR	L-SVM	NNet	RF	GPR	L-SVM	NNet
Areas under the receiver operating characteristic curve								
EN3	98.3	98.1	84.7	94.6	67.3	72.1	75.2	73.3
EN6	98.1	98.3	89.8	94.0	74.1	80.6	82.4	82.5
F_1-scores								
EN3	90.0	89.9	79.4	90.5	60.9	64.7	65.4	63.6
EN6	93.0	93.0	84.7	91.6	63.8	66.9	67.4	69.5

Table 3. The performance of our best-performing natural language processing (NLP)–radiomics pipeline (neural networks with the ensemble 6 region of interest) on the training and test sets using NLP and manually extracted pain labels, together with the results from a prior study by Wakabayashi et al [13].

	Accuracy	Precision	Sensitivity	Specificity	F_1 -score	AUC ^a
This study (training set)	92.4	93.2	92.4	86.4	91.6	94.0
This study (test set)	81.0	67.9	59.2	85.3	69.5	82.5
This study (training set); using manual pain scores	94.2	94.8	98.7	89.7	94.4	98.1
This study (test set); using manual pain scores	83.5	64.9	64.7	85.7	68.0	82.3
Wakabayashi et al [13] (training test only)	73.9	— ^b	71.0	86.0	—	82.0

^aAUC: area under the receiver operating characteristic curve.

^bNot determined.

Discussion

Underestimation and undertreatment of cancer pain can significantly diminish the quality of life of patients with cancer. Accordingly, systems that can objectively measure cancer pain

have the potential to improve quality of life. In this study, we created a scalable NLP-radiomics pain identification pipeline. Our pipeline is designed for palliative treatment for patients with cancer undergoing radiotherapy therapy, for whom there are typically just 2 contemporaneous sources of relevant medical

information at the time of the treatment: consultation notes and simulation CT images. We used an NLP pipeline to extract physician-reported pain scores from radiotherapy consultation notes. NLP-extracted pain scores are appropriate, when structured patient-reported pain scores are unavailable (as is the case for at least 25% to 35% of all patients with cancer [13,48] and for all patients with cancer receiving palliative care who are treated with radiotherapy at our institution at the time the data were used in this study). Our lesion center point-based spherical ROI delineation method significantly sped up the ROI segmentation procedure, enabling us to rapidly delineate BM center points in 176 images in this study. For comparison, the radiomics pipeline that was developed by Wakabayashi et al [13] required full 3D segmentation of each ROI (69 images).

Due to the unbalanced nature of BM pain, our data set contained significantly fewer “no pain” samples. In order to better train our models, we applied SMOTE resampling to the training set to balance the number of samples with the NLP-extracted “pain” and “no pain” labels. We did not apply any resampling techniques to our test (hold out) set to maintain the original sample imbalance. Therefore, while our training set was balanced, our test set had 5 times more “pain” cases than “no pain” cases (136 pain versus 27 no pain cases). This caused a significant change in the pipeline’s performance between the training and test sets. It has been shown that oversampling improves the overall performance of machine learning models, but the effect is stronger on the training set due to the inclusion of replicated samples in the cross-validation subsets [49]. Moreover, the imbalance in our test set led to high specificity (ability to properly identify pain instances) and low sensitivity (ability to correctly identify no pain cases) in the performance evaluation. For comparison, the sample imbalance reported by Wakabayashi et al [13] was 2:1, resulting in a more balanced relationship between the sensitivity and specificity of their model.

The performance of our pipeline did not improve much when we trained and tested it using expert-extracted pain labels (best-available ground-truth). This might be the case because, in the first experiment, we both trained and tested our pipeline using NLP-extracted pain labels, and in the second experiment, we both trained and tested our pipeline using expert-extracted pain labels. Consequently, after being trained with one set of labels (NLP- or expert-extracted), our pipeline performed well on the test set that was labeled using the same method (NLP or expert). We also demonstrated that our pipeline’s performance is comparable to that of Wakabayashi et al [13], who achieved their results using patient-reported pain labels.

Our pipeline performed significantly better on the EN6 ROIs than on the EN3 ROIs. This could be the case because in comparison to EN3, our EN6 ROIs include additional ROIs with sizes of 20, 30, and 50 mm. From visual inspection, we suspect that, in addition to the characteristics of the BM lesion itself, its location (eg, its proximity to the spinal cord) may be a significant contributor to the BM pain. As a result, larger ROIs enable our algorithm to extract characteristics from outside the BM lesion. Wakabayashi et al [13] also demonstrated the effectiveness of using ROIs outside of the BM lesion.

We are unable to offer a convincing explanation as to why neural networks outperformed random forest and support vector machine classifiers in our analysis. Notwithstanding, it has been demonstrated that neural network classifiers perform better when applied to more difficult problems and larger data sets, while random forest and support vector machine classifiers typically perform well with smaller data sets [46,50,51].

Our pipeline was successful in extracting radiomics biomarkers capable of distinguishing between painful and painless BM lesions. These biomarkers potentially provide the opportunity to objectively identify clinical pain-related indicators that may aid in the diagnosis, treatment, and understanding of BM pain.

Our work has several limitations. First, we used data from a single center for this retrospective study. A multicenter study with a larger data set is necessary to assess the generalizability of our radiomics pipeline for pain quantification. We anticipate that the performance of our NLP-radiomics pipeline will vary based on the pain scoring systems of the cohorts tested. Second, by using lesion center point-based geometrical ROIs, we ignored lesion characteristics such as size and shape, which may be important in the context of pain. Although we used Hounsfield units intensity thresholding to preserve some tumor information, we are considering implementing deep learning-based ROI segmentation in the future as it may better account for full tumor and surrounding tissue characteristics. Lastly, we used SMOTE resampling to address the issue of class imbalance. An alternative solution might be to develop cost-sensitive machine learning classifiers that account for the cost of misclassifying minority samples [52]. However, there is no clear consensus in the literature on whether cost-sensitive learning outperforms resampling [53]. A model that can differentiate between painful and painless lesions from medical imaging is a critical component of any possible radiomics-based pain quantification pipeline. This work not only shows the feasibility of developing a pain quantification tool, but also it removes some of the barriers to its development. As a result, our future work will be to apply our pipeline to patients’ past and current CT images and consultation notes in order to develop a longitudinal model of pain. Such a model should take into account not only images (taken before, during, and after delivering radiotherapy) but also other internal and external parameters that can influence how pain evolves over time (such as primary cancer type, radiation dose, other treatments, and pain medications). Also, it will include patient-reported pain scores to provide more accurate ground-truth pain labels in order to develop a more robust deep learning-based NLP pipeline [24,54]. This, however, is beyond the scope of this investigation.

In conclusion, we demonstrated that our NLP and radiomics-based machine learning pipeline can effectively differentiate between painful and painless BM lesions in simulation CT images using ensemble lesion center point-based geometrical ROIs. Using NLP-extracted pain labels in conjunction with lesion center point-based radiomics features is time efficient. This helps to pave the way for the development of quickly trained and efficient clinical AI-based decision-making tools that can objectively measure cancer pain. Such a tool may help alleviate the burden of pain management and improve the quality of life of patients with BMs.

Acknowledgments

This research was supported by the startup grant of J Kildea at the Research Institute of the McGill University Health Centre (RI-MUHC), the Ruth and Alex Dworkin scholarship award from the Faculty of Medicine and Health Sciences at McGill University, an RI-MUHC studentship award, a Grad Excellence Award-00293 from the Department of Physics at McGill University, Fonds de recherche du Québec - Santé (FRQS), and by the CREATE Responsible Health and Healthcare Data Science (SDRDS) grant of the Natural Sciences and Engineering Research Council. The authors would like to thank Dr Luc Galarneau for his help with statistical analysis.

Conflicts of Interest

None declared.

Multimedia Appendix 1

Sample data.

[[DOCX File , 103 KB - ai_v2i1e44779_app1.docx](#)]

References

- van den Beuken-van Everdingen MH, Hochstenbach LM, Joosten EA, Tjan-Heijnen VC, Janssen DJ. Update on prevalence of pain in patients with cancer: systematic review and meta-analysis. *J Pain Symptom Manage* 2016 Jun;51(6):1070-1090.e9 [FREE Full text] [doi: [10.1016/j.jpainsymman.2015.12.340](https://doi.org/10.1016/j.jpainsymman.2015.12.340)] [Medline: [27112310](https://pubmed.ncbi.nlm.nih.gov/27112310/)]
- McQuay HJ, Collins SL, Carroll D, Moore RA, Derry S. WITHDRAWN: radiotherapy for the palliation of painful bone metastases. *Cochrane Database Syst Rev* 2013 Nov 22;2013(11):CD001793 [FREE Full text] [doi: [10.1002/14651858.CD001793.pub2](https://doi.org/10.1002/14651858.CD001793.pub2)] [Medline: [24271498](https://pubmed.ncbi.nlm.nih.gov/24271498/)]
- Grossman SA. Undertreatment of cancer pain: barriers and remedies. *Support Care Cancer* 1993 Mar;1(2):74-78. [doi: [10.1007/bf00366899](https://doi.org/10.1007/bf00366899)]
- Cleeland CS, Janjan NA, Scott CB, Seiferheld WF, Curran WJ. Cancer pain management by radiotherapists: a survey of radiation therapy oncology group physicians. *Int J Radiat Oncol Biol Phys* 2000 Apr 01;47(1):203-208. [doi: [10.1016/s0360-3016\(99\)00276-x](https://doi.org/10.1016/s0360-3016(99)00276-x)] [Medline: [10758325](https://pubmed.ncbi.nlm.nih.gov/10758325/)]
- Tracey I, Woolf CJ, Andrews NA. Composite pain biomarker signatures for objective assessment and effective treatment. *Neuron* 2019 Mar 06;101(5):783-800 [FREE Full text] [doi: [10.1016/j.neuron.2019.02.019](https://doi.org/10.1016/j.neuron.2019.02.019)] [Medline: [30844399](https://pubmed.ncbi.nlm.nih.gov/30844399/)]
- Xu X, Huang Y. Objective pain assessment: a key for the management of chronic pain. *F1000Res* 2020 Jan 23;9:35 [FREE Full text] [doi: [10.12688/f1000research.20441.1](https://doi.org/10.12688/f1000research.20441.1)] [Medline: [32047606](https://pubmed.ncbi.nlm.nih.gov/32047606/)]
- Niculescu AB, Le-Niculescu H, Levey DF, Roseberry K, Soe KC, Rogers J, et al. Towards precision medicine for pain: diagnostic biomarkers and repurposed drugs. *Mol Psychiatry* 2019 Apr;24(4):501-522 [FREE Full text] [doi: [10.1038/s41380-018-0345-5](https://doi.org/10.1038/s41380-018-0345-5)] [Medline: [30755720](https://pubmed.ncbi.nlm.nih.gov/30755720/)]
- Diaz MM, Caylor J, Strigo I, Lerman I, Henry B, Lopez E, et al. Toward composite pain biomarkers of neuropathic pain-focus on peripheral neuropathic pain. *Front Pain Res (Lausanne)* 2022 May 11;3:869215 [FREE Full text] [doi: [10.3389/fpain.2022.869215](https://doi.org/10.3389/fpain.2022.869215)] [Medline: [35634449](https://pubmed.ncbi.nlm.nih.gov/35634449/)]
- Furfari A, Wan BA, Ding K, Wong A, Zhu L, Bezjak A, et al. Genetic biomarkers associated with pain flare and dexamethasone response following palliative radiotherapy in patients with painful bone metastases. *Ann Palliat Med* 2017 Dec;6(Suppl 2):S240-S247. [doi: [10.21037/apm.2017.09.04](https://doi.org/10.21037/apm.2017.09.04)] [Medline: [29156912](https://pubmed.ncbi.nlm.nih.gov/29156912/)]
- Gunn J, Hill MM, Cotten BM, Deer TR. An analysis of biomarkers in patients with chronic pain. *Pain Physician* 2020 Jan;23(1):E41-E49 [FREE Full text] [Medline: [32013287](https://pubmed.ncbi.nlm.nih.gov/32013287/)]
- Marchi A, Vellucci R, Mameli S, Rita Piredda A, Finco G. Pain biomarkers. *Clinical Drug Investigation* 2009;29(Supplement 1):41-46. [doi: [10.2165/0044011-200929001-00006](https://doi.org/10.2165/0044011-200929001-00006)]
- Ota Y, Connolly M, Srinivasan A, Kim J, Capizzano AA, Moritani T. Mechanisms and origins of spinal pain: from molecules to anatomy, with diagnostic clues and imaging findings. *Radiographics* 2020 Jul;40(4):1163-1181. [doi: [10.1148/rg.2020190185](https://doi.org/10.1148/rg.2020190185)] [Medline: [32501739](https://pubmed.ncbi.nlm.nih.gov/32501739/)]
- Wakabayashi K, Koide Y, Aoyama T, Shimizu H, Miyauchi R, Tanaka H, et al. A predictive model for pain response following radiotherapy for treatment of spinal metastases. *Sci Rep* 2021 Jun 18;11(1):12908 [FREE Full text] [doi: [10.1038/s41598-021-92363-0](https://doi.org/10.1038/s41598-021-92363-0)] [Medline: [34145367](https://pubmed.ncbi.nlm.nih.gov/34145367/)]
- Carlson LA, Hooten WM. Pain-linguistics and natural language processing. *Mayo Clin Proc Innov Qual Outcomes* 2020 Jun;4(3):346-347 [FREE Full text] [doi: [10.1016/j.mayocpiqo.2020.01.005](https://doi.org/10.1016/j.mayocpiqo.2020.01.005)] [Medline: [32542226](https://pubmed.ncbi.nlm.nih.gov/32542226/)]
- Dave AD, Ruano G, Kost J, Wang X. Automated extraction of pain symptoms: a natural language approach using electronic health records. *Pain Physician* 2022 Mar;25(2):E245-E254. [Medline: [35322976](https://pubmed.ncbi.nlm.nih.gov/35322976/)]

16. Tighe PJ, Sannapaneni B, Fillingim RB, Doyle C, Kent M, Shickel B, et al. Forty-two million ways to describe pain: topic modeling of 200,000 PubMed pain-related abstracts using natural language processing and deep learning-based text generation. *Pain Med* 2020 Nov 01;21(11):3133-3160 [FREE Full text] [doi: [10.1093/pm/pnaa061](https://doi.org/10.1093/pm/pnaa061)] [Medline: [32249306](https://pubmed.ncbi.nlm.nih.gov/32249306/)]
17. Matsangidou M, Liampas A, Pittara M, Pattichi CS, Zis P. Machine learning in pain medicine: an up-to-date systematic review. *Pain Ther* 2021 Dec 26;10(2):1067-1084 [FREE Full text] [doi: [10.1007/s40122-021-00324-2](https://doi.org/10.1007/s40122-021-00324-2)] [Medline: [34568998](https://pubmed.ncbi.nlm.nih.gov/34568998/)]
18. Neijenhuijs KI, Peeters CFW, van Weert H, Cuijpers P, Leeuw IV. Symptom clusters among cancer survivors: what can machine learning techniques tell us? *BMC Med Res Methodol* 2021 Aug 16;21(1):166 [FREE Full text] [doi: [10.1186/s12874-021-01352-4](https://doi.org/10.1186/s12874-021-01352-4)] [Medline: [34399698](https://pubmed.ncbi.nlm.nih.gov/34399698/)]
19. Hong JH, Jung J, Jo A, Nam Y, Pak S, Lee S, et al. Development and validation of a radiomics model for differentiating bone islands and osteoblastic bone metastases at abdominal CT. *Radiology* 2021 Jun;299(3):626-632. [doi: [10.1148/radiol.2021203783](https://doi.org/10.1148/radiol.2021203783)] [Medline: [33787335](https://pubmed.ncbi.nlm.nih.gov/33787335/)]
20. Sun W, Liu S, Guo J, Liu S, Hao D, Hou F, et al. A CT-based radiomics nomogram for distinguishing between benign and malignant bone tumours. *Cancer Imaging* 2021 Feb 06;21(1):20 [FREE Full text] [doi: [10.1186/s40644-021-00387-6](https://doi.org/10.1186/s40644-021-00387-6)] [Medline: [33549151](https://pubmed.ncbi.nlm.nih.gov/33549151/)]
21. Naseri H, Skamene S, Tolba M, Faye MD, Ramia P, Khriuguian J, et al. Radiomics-based machine learning models to distinguish between metastatic and healthy bone using lesion-center-based geometric regions of interest. *Sci Rep* 2022 Jun 14;12(1):9866 [FREE Full text] [doi: [10.1038/s41598-022-13379-8](https://doi.org/10.1038/s41598-022-13379-8)] [Medline: [35701461](https://pubmed.ncbi.nlm.nih.gov/35701461/)]
22. Mashayekhi R, Parekh VS, Faghih M, Singh VK, Jacobs MA, Zaheer A. Radiomic features of the pancreas on CT imaging accurately differentiate functional abdominal pain, recurrent acute pancreatitis, and chronic pancreatitis. *Eur J Radiol* 2020 Feb;123:108778 [FREE Full text] [doi: [10.1016/j.ejrad.2019.108778](https://doi.org/10.1016/j.ejrad.2019.108778)] [Medline: [31846864](https://pubmed.ncbi.nlm.nih.gov/31846864/)]
23. Vedantam A, Hassan I, Kotrotsou A, Hassan A, Zinn PO, Viswanathan A, et al. Magnetic resonance-based radiomic analysis of radiofrequency lesion predicts outcomes after percutaneous cordotomy: a feasibility study. *Oper Neurosurg (Hagerstown)* 2020 Jun 01;18(6):721-727. [doi: [10.1093/ons/ozp288](https://doi.org/10.1093/ons/ozp288)] [Medline: [31665446](https://pubmed.ncbi.nlm.nih.gov/31665446/)]
24. Devlin J, Chang MW, Lee K, Toutanova K. BERT: pre-training of deep bidirectional transformers for language understanding. arXiv Preprint posted online October 11, 2018. [doi: [10.48550/arXiv.1810.04805](https://doi.org/10.48550/arXiv.1810.04805)]
25. Naseri H, Kafi K, Skamene S, Tolba M, Faye MD, Ramia P, et al. Development of a generalizable natural language processing pipeline to extract physician-reported pain from clinical reports: Generated using publicly-available datasets and tested on institutional clinical reports for cancer patients with bone metastases. *J Biomed Inform* 2021 Aug;120:103864 [FREE Full text] [doi: [10.1016/j.jbi.2021.103864](https://doi.org/10.1016/j.jbi.2021.103864)] [Medline: [34265451](https://pubmed.ncbi.nlm.nih.gov/34265451/)]
26. Elbattah M, Arnaud É, Gignon M, Dequen G. The role of text analytics in healthcare: a review of recent developments and applications. 2021 Presented at: 14th International Joint Conference on Biomedical Engineering Systems and Technologies - Scale-IT-up; February 11-13, 2021; Online. [doi: [10.5220/0010414508250832](https://doi.org/10.5220/0010414508250832)]
27. Smith TMF, Cochran WG. Sampling techniques, second edition. *Applied Statistics* 1964;13(1):54. [doi: [10.2307/2985224](https://doi.org/10.2307/2985224)]
28. Freedman D, Pisani R, Purves R. *Statistics: Fourth International Student Edition*. New York, NY: W.W. Norton & Company; 2007.
29. Aronson AR. Effective mapping of biomedical text to the UMLS Metathesaurus: the MetaMap program. *Proc AMIA Symp* 2001:17-21 [FREE Full text] [Medline: [11825149](https://pubmed.ncbi.nlm.nih.gov/11825149/)]
30. McCray AT, Aronson AR, Browne AC, Rindflesch TC, Razi A, Srinivasan S. UMLS knowledge for biomedical language processing. *Bull Med Libr Assoc* 1993 Apr;81(2):184-194 [FREE Full text] [Medline: [8472004](https://pubmed.ncbi.nlm.nih.gov/8472004/)]
31. hn617/texTRACTOR: texTRACTOR. Zenodo. 2021. URL: <https://zenodo.org/record/4649625> [accessed 2023-04-18]
32. hn617/diCOMBINE: diCOMBINE. Zenodo. 2021. URL: <https://zenodo.org/record/5218743> [accessed 2023-04-18]
33. Busscher I, Ploegmakers JJW, Verkerke GJ, Veldhuizen AG. Comparative anatomical dimensions of the complete human and porcine spine. *Eur Spine J* 2010 Jul 26;19(7):1104-1114 [FREE Full text] [doi: [10.1007/s00586-010-1326-9](https://doi.org/10.1007/s00586-010-1326-9)] [Medline: [20186441](https://pubmed.ncbi.nlm.nih.gov/20186441/)]
34. Deglinc HJ, Rangayyan RM, Ayres FJ, Boag GS, Zuffo MK. Three-dimensional segmentation of the tumor in computed tomographic images of neuroblastoma. *J Digit Imaging* 2006 Aug 25;20(1):72-87. [doi: [10.1007/10278-006-0769-3](https://doi.org/10.1007/10278-006-0769-3)]
35. Ulano A, Bredella MA, Burke P, Chebib I, Simeone FJ, Huang AJ, et al. Distinguishing untreated osteoblastic metastases from enostoses using CT attenuation measurements. *Am J Roentgenol* 2016 Aug;207(2):362-368. [doi: [10.2214/ajr.15.15559](https://doi.org/10.2214/ajr.15.15559)]
36. Smoothing Images. OpenCV. URL: https://docs.opencv.org/4.x/d4/d13/tutorial_py_filtering.html [accessed 2023-09-18]
37. mhe/pynrrd: v0.4.3 Released. Zenodo. 2022. URL: <https://zenodo.org/record/6501810> [accessed 2023-04-18]
38. Nearly Raw Raster Data. URL: <http://teem.sourceforge.net/nrrd/index.html> [accessed 2022-09-04]
39. Lemaitre G, Nogueira F, Aridas CK. Imbalanced-learn: A Python Toolbox to Tackle the Curse of Imbalanced Datasets in Machine Learning. *J Mach Learn Res* 2017;18:1-5 [FREE Full text]
40. Low E. Review of Understanding Basic Statistics. *Am Stat* 1998;52(2):198. [doi: [10.2307/2685480](https://doi.org/10.2307/2685480)]
41. Torvik K, Hølen J, Kaasa S, Kirkevold ?, Holtan A, Kongsgaard U, et al. Pain in elderly hospitalized cancer patients with bone metastases in Norway. *Int J Palliat Nurs* 2008 May;14(5):238-245. [doi: [10.12968/ijpn.2008.14.5.29491](https://doi.org/10.12968/ijpn.2008.14.5.29491)] [Medline: [18563017](https://pubmed.ncbi.nlm.nih.gov/18563017/)]
42. Chawla NV, Bowyer KW, Hall LO, Kegelmeyer WP. SMOTE: synthetic minority over-sampling technique. *J Artif Intell Res* 2002 Jun 01;16:321-357. [doi: [10.1613/jair.953](https://doi.org/10.1613/jair.953)]

43. Xie C, Du R, Ho JW, Pang HH, Chiu KW, Lee EY, et al. Effect of machine learning re-sampling techniques for imbalanced datasets in F-FDG PET-based radiomics model on prognostication performance in cohorts of head and neck cancer patients. *Eur J Nucl Med Mol Imaging* 2020 Nov 06;47(12):2826-2835. [doi: [10.1007/s00259-020-04756-4](https://doi.org/10.1007/s00259-020-04756-4)] [Medline: [32253486](https://pubmed.ncbi.nlm.nih.gov/32253486/)]
44. Tibshirani R. Regression shrinkage and selection via The Lasso: a retrospective. *J R Stat Soc Series B Stat Methodol* 2011;73(3):273-282. [doi: [10.1111/j.1467-9868.2011.00771.x](https://doi.org/10.1111/j.1467-9868.2011.00771.x)]
45. Yin P, Mao N, Chen H, Sun C, Wang S, Liu X, et al. Machine and deep learning based radiomics models for preoperative prediction of benign and malignant sacral tumors. *Front Oncol* 2020 Oct 16;10:564725 [FREE Full text] [doi: [10.3389/fonc.2020.564725](https://doi.org/10.3389/fonc.2020.564725)] [Medline: [33178593](https://pubmed.ncbi.nlm.nih.gov/33178593/)]
46. Shur JD, Doran SJ, Kumar S, Ap Dafydd D, Downey K, O'Connor JPB, et al. Radiomics in oncology: a practical guide. *Radiographics* 2021 Oct;41(6):1717-1732. [doi: [10.1148/rg.2021210037](https://doi.org/10.1148/rg.2021210037)] [Medline: [34597235](https://pubmed.ncbi.nlm.nih.gov/34597235/)]
47. Fawcett T. An introduction to ROC analysis. *Pattern Recognit Lett* 2006 Jun;27(8):861-874. [doi: [10.1016/j.patrec.2005.10.010](https://doi.org/10.1016/j.patrec.2005.10.010)]
48. Fleischman RJ, Frazer DG, Daya M, Jui J, Newgard CD. Effectiveness and safety of fentanyl compared with morphine for out-of-hospital analgesia. *Prehosp Emerg Care* 2010 Mar 03;14(2):167-175 [FREE Full text] [doi: [10.3109/10903120903572301](https://doi.org/10.3109/10903120903572301)] [Medline: [20199230](https://pubmed.ncbi.nlm.nih.gov/20199230/)]
49. Fernández A, García S, Galar M, Prati RC, Krawczyk B, Herrera F. Learning from imbalanced data streams. In: *Learning from Imbalanced Data Sets*. Cham: Springer; 2018.
50. Sun Q, Lin X, Zhao Y, Li L, Yan K, Liang D, et al. Deep learning vs. radiomics for predicting axillary lymph node metastasis of breast cancer using ultrasound images: don't forget the peritumoral region. *Front Oncol* 2020 Jan 31;10:53 [FREE Full text] [doi: [10.3389/fonc.2020.00053](https://doi.org/10.3389/fonc.2020.00053)] [Medline: [32083007](https://pubmed.ncbi.nlm.nih.gov/32083007/)]
51. Lisson CS, Lisson CG, Mezger MF, Wolf D, Schmidt SA, Thaiss WM, et al. Deep neural networks and machine learning radiomics modelling for prediction of relapse in mantle cell lymphoma. *Cancers (Basel)* 2022 Apr 15;14(8):2008 [FREE Full text] [doi: [10.3390/cancers14082008](https://doi.org/10.3390/cancers14082008)] [Medline: [35454914](https://pubmed.ncbi.nlm.nih.gov/35454914/)]
52. Thai-Nghe N, Gantner Z, Schmidt-Thieme L. Cost-sensitive learning methods for imbalanced data. 2010 Presented at: The 2010 International Joint Conference on Neural Networks (IJCNN); July 18-23, 2010; Barcelona. [doi: [10.1109/ijcnn.2010.5596486](https://doi.org/10.1109/ijcnn.2010.5596486)]
53. Liu A, Martin C, La Cour B, Ghosh J. Effects of Oversampling Versus Cost-Sensitive Learning for Bayesian and SVM Classifiers. In: Stahlbock R, Crone S, Lessmann S, editors. *Data Mining. Annals of Information Systems (volume 8)*. Boston, MA: Springer; 2010.
54. Tamang S, Humbert-Droz M, Gianfrancesco M, Izadi Z, Schmajuk G, Yazdany J. Practical considerations for developing clinical natural language processing systems for population health management and measurement. *JMIR Med Inform* 2023 Jan 03;11:e37805 [FREE Full text] [doi: [10.2196/37805](https://doi.org/10.2196/37805)] [Medline: [36595345](https://pubmed.ncbi.nlm.nih.gov/36595345/)]

Abbreviations

- AI:** artificial intelligence
API: average pain intensity
AUC: area under the receiver operating characteristic curve
BM: bone metastasis
CT: computed tomography
EN: ensemble
NLP: natural language processing
ROC: receiver operating characteristic
ROI: region of interest
SMOTE: Synthetic Minority Oversampling Technique
SP: spherical
VDP: verbally declared pain

Edited by K El Emam, B Malin; submitted 02.12.22; peer-reviewed by E Hmouda, SY Wang, M Elbattah; comments to author 06.02.23; revised version received 12.03.23; accepted 01.04.23; published 22.05.23.

Please cite as:

Naseri H, Skamene S, Tolba M, Faye MD, Ramia P, Khriouan J, David M, Kildea J
A Scalable Radiomics- and Natural Language Processing-Based Machine Learning Pipeline to Distinguish Between Painful and Painless Thoracic Spinal Bone Metastases: Retrospective Algorithm Development and Validation Study
JMIR AI 2023;2:e44779
 URL: <https://ai.jmir.org/2023/1/e44779>
 doi: [10.2196/44779](https://doi.org/10.2196/44779)
 PMID: [38875572](https://pubmed.ncbi.nlm.nih.gov/38875572/)

©Hossein Naseri, Sonia Skamene, Marwan Tolba, Mame Daro Faye, Paul Ramia, Julia Khriouian, Marc David, John Kildea. Originally published in JMIR AI (<https://ai.jmir.org>), 22.05.2023. This is an open-access article distributed under the terms of the Creative Commons Attribution License (<https://creativecommons.org/licenses/by/4.0/>), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR AI, is properly cited. The complete bibliographic information, a link to the original publication on <https://www.ai.jmir.org/>, as well as this copyright and license information must be included.

Original Paper

Detecting Ground Glass Opacity Features in Patients With Lung Cancer: Automated Extraction and Longitudinal Analysis via Deep Learning–Based Natural Language Processing

Kyeryoung Lee^{1*}, PhD; Zongzhi Liu^{1*}, PhD; Urmila Chandran², PhD; Iftekhar Kalsekar², PhD; Balaji Laxmanan², MD; Mitchell K Higashi¹, PhD; Tomi Jun¹, MD; Meng Ma¹, PhD; Minghao Li¹, MSc; Yun Mai¹, PhD; Christopher Gilman¹, BSc; Tongyu Wang¹, BSc; Lei Ai¹, PhD; Parag Aggarwal¹, PhD; Qi Pan¹, PhD; William Oh³, MD; Gustavo Stolovitzky¹, PhD; Eric Schadt³, PhD; Xiaoyan Wang¹, PhD

¹Sema4, Stamford, CT, United States

²Lung Cancer Initiative, Johnson & Johnson, New Brunswick, NJ, United States

³Icahn School of Medicine at Mount Sinai, New York, NY, United States

*these authors contributed equally

Corresponding Author:

Xiaoyan Wang, PhD

Sema4

333 Ludlow Street

Stamford, CT, 06902

United States

Phone: 1 800 298 6470

Email: xw108@caa.columbia.edu

Abstract

Background: Ground-glass opacities (GGOs) appearing in computed tomography (CT) scans may indicate potential lung malignancy. Proper management of GGOs based on their features can prevent the development of lung cancer. Electronic health records are rich sources of information on GGO nodules and their granular features, but most of the valuable information is embedded in unstructured clinical notes.

Objective: We aimed to develop, test, and validate a deep learning–based natural language processing (NLP) tool that automatically extracts GGO features to inform the longitudinal trajectory of GGO status from large-scale radiology notes.

Methods: We developed a bidirectional long short-term memory with a conditional random field–based deep-learning NLP pipeline to extract GGO and granular features of GGO retrospectively from radiology notes of 13,216 lung cancer patients. We evaluated the pipeline with quality assessments and analyzed cohort characterization of the distribution of nodule features longitudinally to assess changes in size and solidity over time.

Results: Our NLP pipeline built on the GGO ontology we developed achieved between 95% and 100% precision, 89% and 100% recall, and 92% and 100% F_1 -scores on different GGO features. We deployed this GGO NLP model to extract and structure comprehensive characteristics of GGOs from 29,496 radiology notes of 4521 lung cancer patients. Longitudinal analysis revealed that size increased in 16.8% (240/1424) of patients, decreased in 14.6% (208/1424), and remained unchanged in 68.5% (976/1424) in their last note compared to the first note. Among 1127 patients who had longitudinal radiology notes of GGO status, 815 (72.3%) were reported to have stable status, and 259 (23%) had increased/progressed status in the subsequent notes.

Conclusions: Our deep learning–based NLP pipeline can automatically extract granular GGO features at scale from electronic health records when this information is documented in radiology notes and help inform the natural history of GGO. This will open the way for a new paradigm in lung cancer prevention and early detection.

(JMIR AI 2023;2:e44537) doi:[10.2196/44537](https://doi.org/10.2196/44537)

KEYWORDS

natural language processing; ground glass opacity; real world data; radiology notes; longitudinal analysis; deep learning; bidirectional long short-term memory (Bi-LSTM); conditional random fields (CRF)

Introduction

The goal of lung cancer treatment is primary prevention, early prediction, and detection of lung malignancy to reduce lung cancer mortality. Currently, prevention screening programs have proven to be effective in the early detection of many cancers [1]. Low-dose computed tomography (CT) has been a standard method for lung cancer screening in the United States since the National Lung Screening Trial in 2011 [2,3]. With the increased utilization of CT scans and advances in CT techniques, the detection rate of pulmonary nodules has increased during the last decade [4]. Approximately 20% to 30% of CT images detect pulmonary nodules with ground-glass opacity (GGO), a subtype of pulmonary nodules [5-7]. GGOs, either pure GGOs (without a solid component) or part-solid GGOs (with a solid component), have gained significant attention in recent years due to their malignancy potential [8-11] ever since Jang and colleagues [12] found that ground-glass attenuation could be a sign of lung adenocarcinoma. However, identifying malignant lesions based on GGO images from CT scans remains a challenge since both benign and malignant lung lesions can appear as GGOs [13-15]. Persistent GGOs, which have not been resolved in subsequent CT scans between 6 and 12 months, are more likely to be associated with precancerous or cancerous conditions, while transient and self-resolving GGOs are benign [16-19]. Other GGO features such as larger baseline nodule size, spiculated shape, upper lobe location, presence of a solid component, and less than 5 nodules in quantity are known to be highly associated with the probability of malignancy [20-23]. Understanding the characteristics and prognosis of GGOs is critical for predicting and preventing lung cancer development by adopting proper management [24,25].

Radiomics is a study field leveraging artificial intelligence (AI) to extract medical information from radiology images. Recent advances in radiomics have significantly improved the accuracy of identifying malignant lesions [26-28] and made possible differentiating etiologies of GGOs [29]. However, limited access to scans, the high cost, and the complexity of processes [30-32] have hindered the routine knowledge extraction from CT scans and prompted the use of patient electronic health records (EHRs). EHRs are rich sources of patients' clinical information including radiological findings [33,34], which are generally captured in unstructured data fields. However, large-scale extraction of GGO information from an enormous collection of unstructured EHR data is almost impossible without leveraging the power of natural language processing (NLP).

NLP is an AI approach that enables extracting large-scale information automatically from clinical notes and presenting the extracted information in a computer interoperable structured format. Over the last 2 decades, NLP has played a critical role in representing medical information that is embedded in

unstructured clinical notes [35-39] and has been applied to the field of radiology [40]. Pons et al [34] systematically reviewed 67 NLP studies in radiology reports and demonstrated how radiology fields benefit from NLP techniques. Linna and Kahn [41] also highlighted the potential benefits of NLP technology in multiple areas, such as improved diagnostic decision-making, patient care, and delivery. Although the development of deep learning methods and transformer models like Bidirectional Encoder Representations From Transformers (BERT) showed a significantly improved impact in named entity recognition and relation extraction [42], these state-of-the-art NLP methods have not been applied yet to extract data on GGOs and their related features. A few shallow NLP parsers have been developed to identify cohorts with GGOs [14,43-46]. Recently, a rule-based GGO NLP algorithm was developed and applied in combination with negation and temporal algorithms to extract and characterize all GGO attributes from radiology reports [4].

This study aimed to investigate the feasibility of developing a deep learning-based NLP model to extract GGO features systematically from radiology notes for the longitudinal analysis of patient-level GGO features on a large scale with ontology-guided contextual embedding and temporal reasoning. The utility of the NLP was then evaluated by deploying it to longitudinal data to assess changes in GGO features longitudinally, which is vital for understanding the natural history of GGOs in the real-world lung cancer setting.

Methods

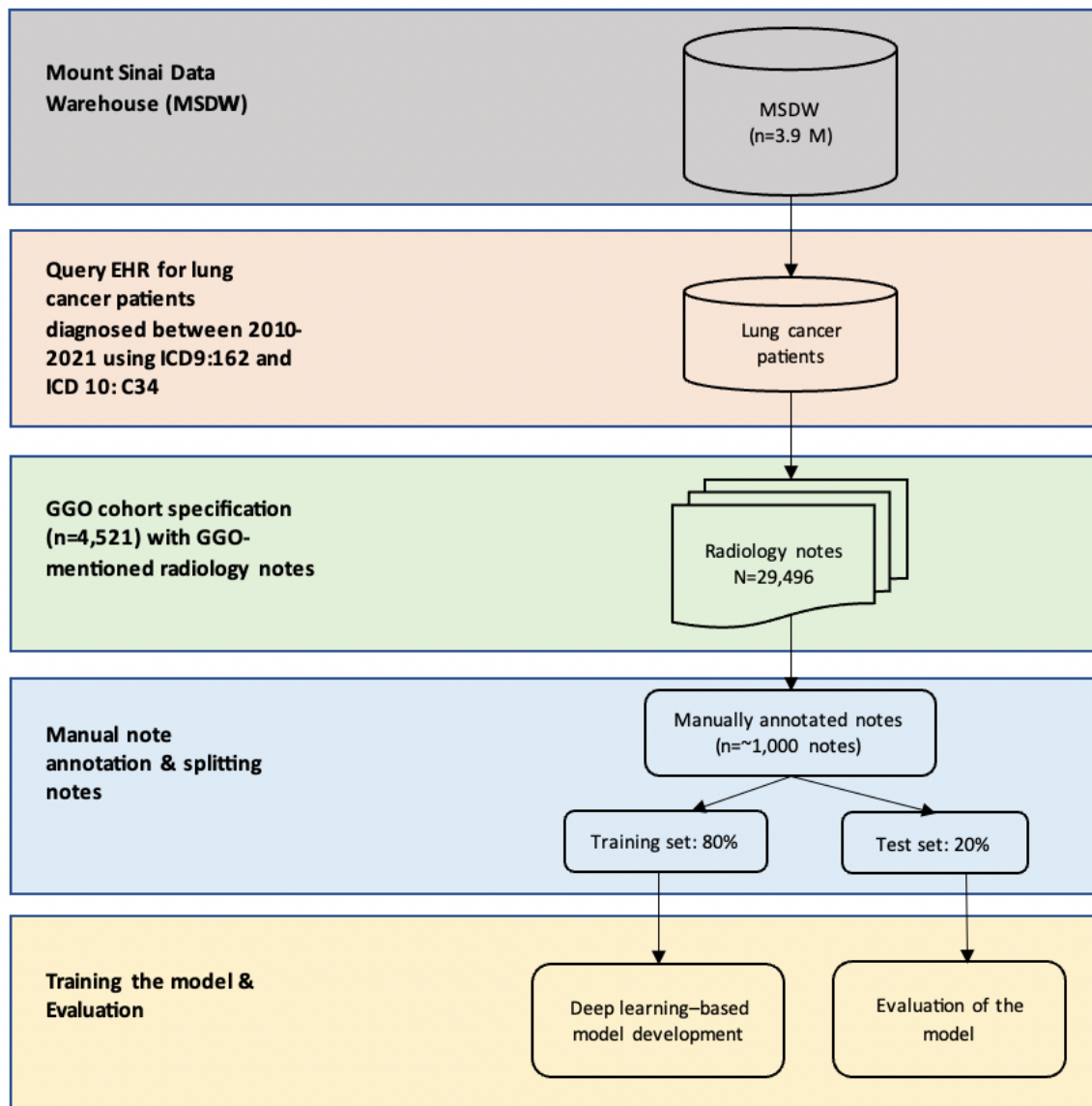
Ethics Approval

This study was approved by the Program for the Protection of Human Subjects at the Mount Sinai School of Medicine (IRB-17-01245).

Study Cohort

The cohort of patients diagnosed with lung cancer between 2010 and 2021 (13,216 patients) was curated from the Mount Sinai/Sema4 Healthcare system, which contains longitudinal data for approximately 3.9 million patients. Demographic and other clinical variables were obtained by either extracting from structured data or curating the relevant information from unstructured clinical notes (ie, radiology notes and progress notes). The study cohort includes (1) pathology-confirmed patients with lung cancer; (2) non-pathology-confirmed patients with lung cancer via ≥ 3 visits and International Classification of Diseases (ICD) lung cancer codes (ICD-9: 162 and ICD-10: C34); and (3) non-pathology-confirmed patients who had < 3 visits with lung cancer ICD codes. We curated these initial lung cancer cohorts to develop and test the GGO NLP pipeline, which can then be applied to other relevant cohorts in the future. [Figure 1](#) shows how we selected study cohorts and their radiology notes from EHRs for the next steps of model training and evaluation.

Figure 1. The workflow of the ground-glass opacity (GGO) natural language (NLP) pipeline. The workflow shows how we selected study cohorts and their radiology notes from EHRs for the next steps of model training and evaluation. EHR: electronic health record; ICD: International Classification of Diseases.



NLP Framework

Overview

The framework we propose to curate GGOs and their related attributes are described as follows: (1) preprocessing and query expansion; (2) GGO ontology construction and annotation; (3) NLP model development; (4) postprocessing and entity normalization; and (5) NLP pipeline evaluation. These are discussed in greater detail in the following subsections.

Preprocessing and Query Expansion

The preprocessing phase focused on query expansion. An initial list of seed terms was obtained from a manual survey of the literature and a review of clinical notes by a clinical researcher and a domain expert (authors KL and MM). A bigram word2vec algorithm [47] was developed to identify additional significant terms potentially related to GGO to ensure the encapsulation of an expansive cohort. The expanded list of query terms was then applied to extract a comprehensive set of GGO-specific

patient notes that were subsequently leveraged for NLP modeling.

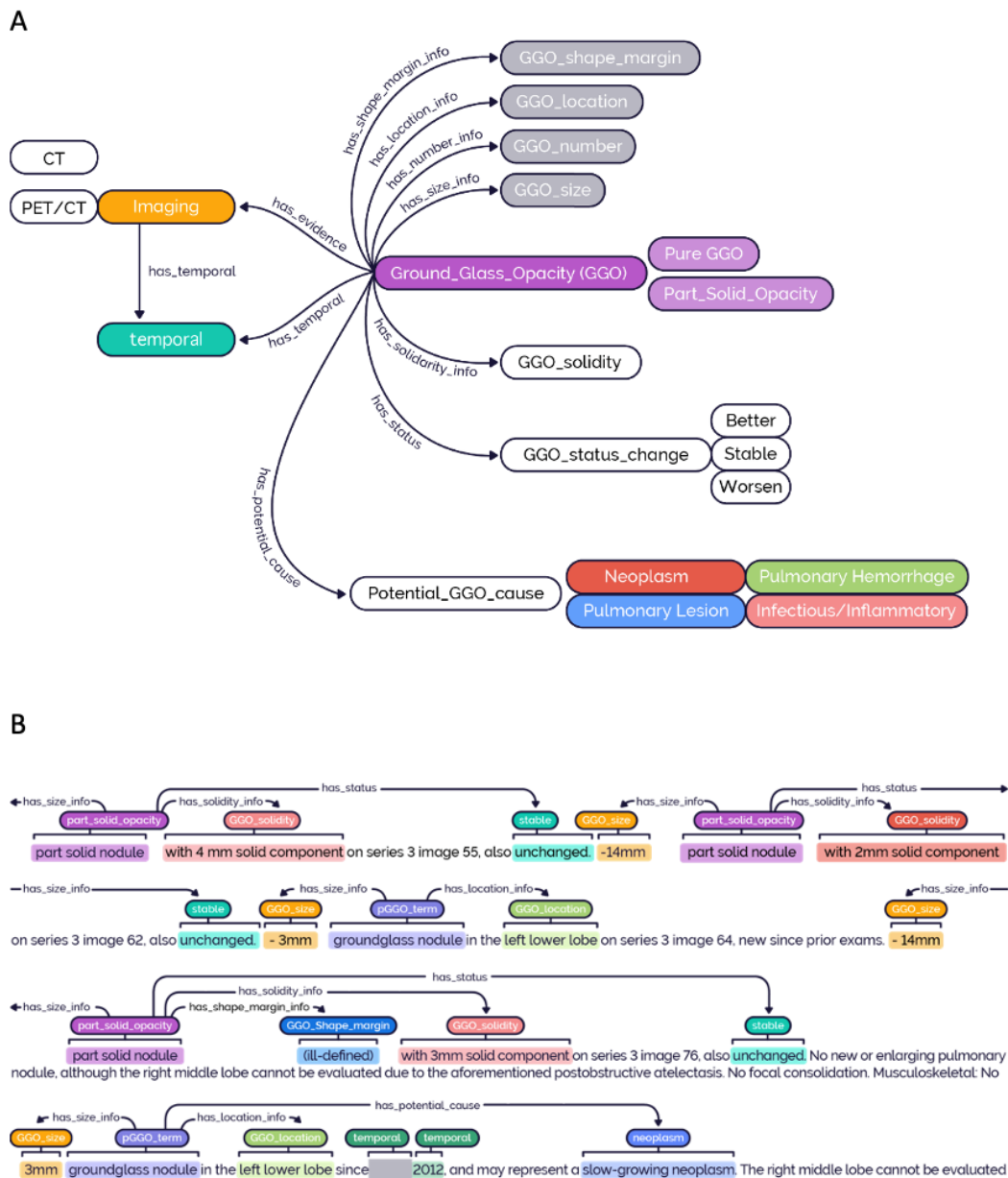
GGO Ontology and Annotation

NLP is the process of simulating an expert's knowledge and understanding of the free text using modeling. As the first step of NLP, we built up an ontology that was established based on clinical expert opinion, comprehensive literature, and patient note review. The GGO ontology includes entities that are critical for cancer prediction based on previous studies and available from our radiology notes. Our GGO ontology includes 15 entities comprising pure GGO, part-solid GGO, GGO size, GGO quantity (number), GGO location, GGO shape/margin, GGO solidity, temporal (date), potential GGO cause (neoplasm, infectious/inflammation, hemorrhage, and other pulmonary lesions), and GGO status change (better, stable, and worsen). Moreover, it has 7 semantic relations between entities: has size information (info), has number info, has location info, has shape/margin info, has solidity info, has status, has a potential

cause (Figure 2A). This ontology was used as a guideline for manual annotation. GGO status change indicates any description of size or solidity changes (eg, increased, getting smaller, getting denser). The primary GGO entities, either pure or part solid, were associated with their attributes like size, location, and so on. Then, 2 independent domain experts manually annotated the 15 entities and 7 semantic relations in the clinical notes (Figure 2B) using the Clinical Language Annotation, Modeling, and Processing (CLAMP) NLP toolkit [48], and a third domain expert (KL) reviewed the annotations.

Since a biomedical concept could be described in heterogeneous forms, continuous discussions and agreement between annotators and domain experts were needed to confirm that the annotations represented the expert’s understanding of biomedical knowledge. Interannotator agreement scores (kappa scores) were measured between the first 2 annotators in the same set of notes until they reached over 90% in entities and over 80% in relation annotation before commencing the independent annotation.

Figure 2. The ontology of ground-glass opacity (GGO) and the sample note with GGO annotations. A) The ontology of GGO. A total of 15 entities and 7 semantic relation types were defined in the GGO ontology. Entity semantic types: GGO location, GGO number, GGO shape/margin, GGO size, GGO solidarity, GGO status change: better, GGO status change: stable, GGO status change: worsen, GGO term: pure GGO term, GGO term: part-solid GGO, potential GGO cause: infectious/inflammatory, potential GGO cause: neoplasm, potential GGO cause: hemorrhage, potential GGO cause: other pulmonary lesions, and temporal. Relation semantic types: has location info, has number info, has shape/margin info, has size info, has solidarity info, has status, and has potential causes. B) Sample deidentified radiology reports with GGO annotations. Each part-solid nodule or ground-glass nodule is associated with attributes (such as size, location, status, change, shape, and/or solidity information) and potential etiologies. The upper panel shows a radiology report with multiple GGOs and their attributes; the lower panel shows a GGO and its associated potential etiologies. CT: computed tomography; PET: positron emission tomography.

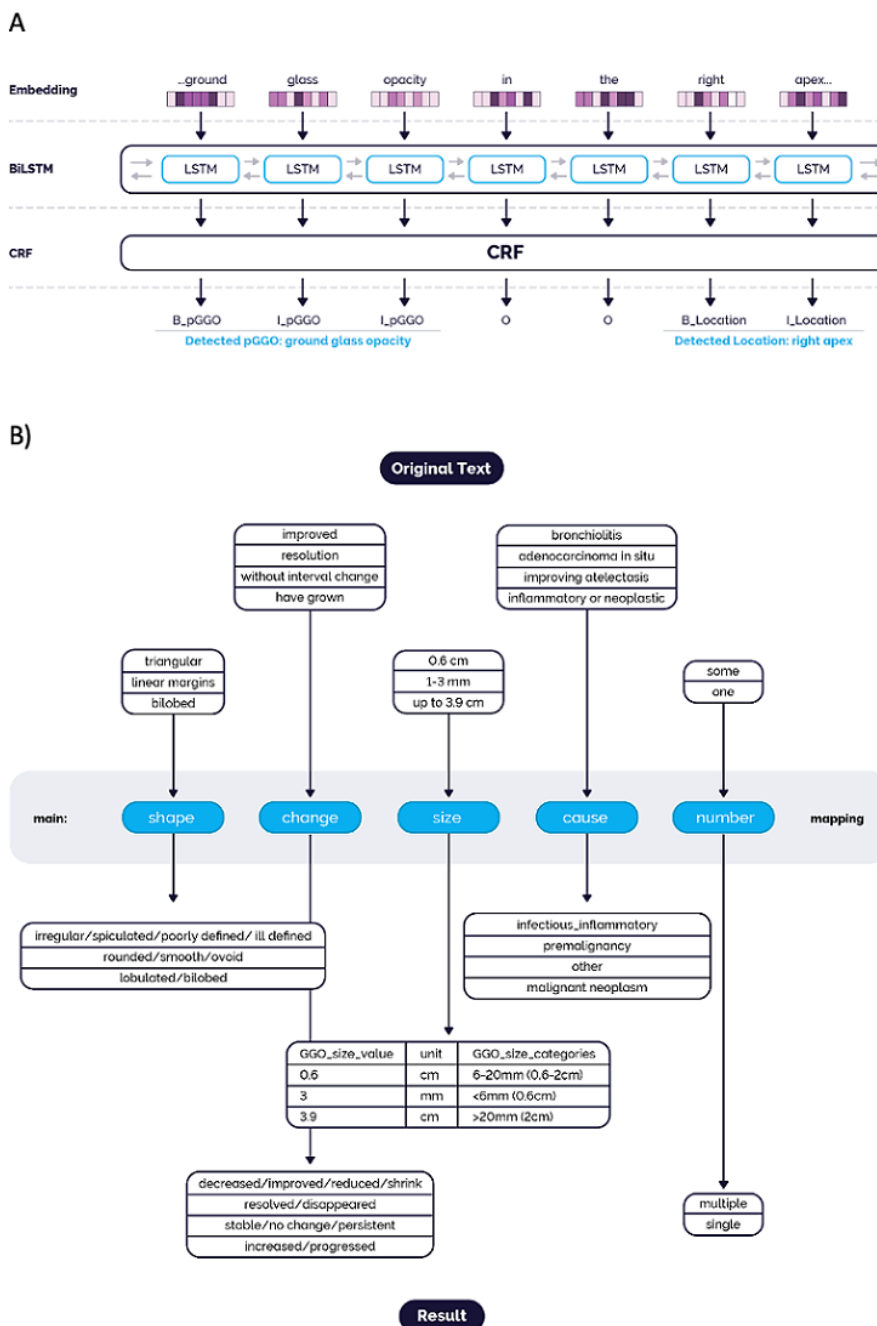


NLP Model Development

A multilayer deep learning architecture was implemented for NLP modeling. The text was first transformed as sequential vectors of characterization in the embedding step. The vectors were then sent to the bidirectional long-short term memory (Bi-LSTM), an artificial neural network of text classification architecture, for pattern recognition in both forward and backward directions [49]. The patterns were sent to the next

layer of a conditional random field (CRF) model to compute prediction probability (Figure 3A) [50]. In the example sentence of Figure 3A, the “ground-glass opacity” is predicated as the entities of “GGO,” while “right apex” is predicated as “location.” The model was trained, calibrated, and tested for optimal performance. Among manually annotated clinical notes, 80% (798/998) were used for training the GGO model and 20% (200/998) were used for validation.

Figure 3. A deep learning natural language processing (NLP) pipeline for ground-glass opacity (GGO) curation and the process of GGO entity normalization. A) Multilayer deep learning NLP architecture for GGO curation. All clinical notes underwent word embedding before being sent to the bidirectional long-short term memory (Bi-LSTM), an artificial neural network of text classification architecture. The outputs were fed to a conditional random fields (CRF) model to predict the GGO entities and relations. B) GGO entity normalization. The raw outputs of NLP models (upper panel) were normalized to standardized concepts (lower panel) for each GGO attribute (middle panel).



Postprocessing and Entity Normalization

A postprocessor was developed to subsequently postcoordinate and refine the output. All predicated entities from the raw text were normalized to standardized concepts based on clinical experts' opinions and were then ready for downstream analysis. Figure 3B illustrates examples of extracted GGO feature entities categorized and normalized for the data analysis. GGO location was extracted and classified into 2 levels; the first level corresponded to a high-level indication of right, left, or bilateral lungs, and the second level corresponded to a more granular indication of the anatomic location like right upper lobe (RUL), right middle lobe (RML), or right lower lobe (RLL), left upper lobe (LUL), and left lower lobe (LLL). We categorized GGO size into 3 groups: <6 mm, 6 to 20 mm, and >20 mm based on expert opinion and the practice guidelines for nonsolid nodules. Potential etiologies found in the notes were classified into 3 subgroups: infectious/inflammatory, malignant, and others, whereby precancerous conditions such as atypical adenomatous hyperplasia and adenocarcinoma in situ were included in the malignant category. Others include all benign pulmonary lesions like fibrosis/scarring and hemorrhage.

NLP Pipeline Evaluation

The performance of the GGO NLP pipeline was estimated in the validation set with precision via the positive predictive value (PPV) and recall via sensitivity, as well as F_1 -score, a balanced score between false positives (FPs) and false negatives (FNs). Recall was calculated as the ratio of the number of entities that were identified by the pipeline over the total number of the corresponding entities in the manually annotated gold standard, such as true positive (TP)/(TP + FN). Precision was measured as the ratio of the number of distinct entities returned by the pipeline that was correct according to the gold standard divided by the total number of entities found by our pipeline, such as TP/(TP + FP). The F_1 -score was calculated as the harmonic mean of PPV and sensitivity, such as $2 \times \text{PPV} \times \text{sensitivity} / (\text{PPV} + \text{sensitivity})$. The manual annotation and training process was repeated with additional manually annotated notes until the model achieved an average F_1 -score >0.8.

Characterization of GGO Cohorts and Longitudinal Analysis of GGOs

To demonstrate the utility of our GGO NLP pipeline, the NLP was deployed to the lung cancer cohort identified in the Mount Sinai/Sema4 data set to identify a cohort of patients with GGOs.

Since the persistence of GGOs is an important indicator of malignancy [18,19], a subset of patients with persistent GGOs was identified by the NLP. Persistence was defined as either patients having multiple GGO reports, except when the last report indicated resolution of the GGO, or patients having only 1 GGO report but with an indication of the increase in the size or quantity or change in solidity. We used the NLP pipeline to identify GGO features from patient notes over time and assessed longitudinal changes in GGO features for this cohort.

To evaluate whether our automatically extracted information was consistent with published findings, such as larger baseline size or upper lobe location of GGOs being highly associated with the malignancy [22], we selected patients who had their first GGO report before lung cancer diagnosis date and performed a descriptive statistical analysis across the natural history of GGOs.

Finally, we extracted patients' demographics and other clinical characteristics including smoking status, comorbidities, and family disease history from structured EHR data to characterize the population with GGOs. All statistical analyses were conducted using R software (R Foundation for Statistical Computing) and done both at the GGO level and patient level depending on the type of assessment.

Results

Patient Characteristics

The distribution of demographic and other clinical characteristics (ie, smoking status, comorbidities, and family history of cancer for the overall GGO cohort) over GGO persistency is shown in Table 1. The average age of the GGO cohort was 68 years; 53.77% (2431/4521) were female, and 52.95% (2394/4521) were White. Smoking data were not available for half the cohort, while among those for whom smoking data were available, 37.63% (1701/4521) of patients were either former or current smokers. Almost 70% (3086/4521) of patients had a history of cancer, and around 13% (606/4521) had a history of chronic obstructive pulmonary disease. The majority (3269/4521, 72.30%) of the GGO cohort had persistent GGOs and similar distributions of patient characteristics as the overall GGO cohort. Most GGO reports were found in the postlung cancer diagnosis period (2815/4251, 62.3%) (Figure S1 in Multimedia Appendix 1).

Table 1. Distribution of demographic and other clinical characterization of GGO^a cohorts.

Variables	Overall (N=4521), n (%)	GGO cohort persistency	
		Persistent GGO (n=3269), n (%)	Nonpersistent GGO (n=1252), n (%)
Gender			
Female	2431 (53.77)	1790 (54.76)	641 (51.20)
Male	2090 (46.23)	1479 (45.24)	611 (48.80)
Race			
White	2394 (52.95)	1700 (52)	694 (55.43)
Other	791 (17.50)	603 (18.45)	188 (15.02)
Black or African American	722 (15.97)	530 (16.21)	192 (15.34)
Unknown	363 (8.03)	244 (7.46)	119 (9.50)
Asian	165 (3.65)	139 (4.25)	26 (2.08)
Native Hawaiian or other Pacific Islander	83 (1.84)	50 (1.53)	33 (2.64)
American Indian or Alaska Native	3 (0.07)	3 (0.09)	0 (0)
Ethnicity			
Not Hispanic or Latino	2442 (54.01)	1864 (57.02)	578 (46.17)
Unknown	1492 (33)	955 (29.21)	537 (42.89)
Hispanic or Latino	519 (11.48)	399 (12.21)	120 (9.58)
Not reported	68 (1.50)	51 (1.56)	17 (1.36)
Smoking status			
No record of smoking	2304 (50.96)	1557 (47.63)	747 (59.66)
Former smoker	1287 (28.47)	996 (30.47)	291 (23.24)
Never smoker	511 (11.30)	395 (12.08)	116 (9.27)
Smoker	414 (9.16)	317 (9.70)	97 (7.75)
Passive smoker	5 (0.11)	4 (0.12)	1 (0.08)
Comorbidities^b			
History of COPD ^c	604 (13.36)	444 (13.58)	160 (12.78)
History of heart disease	1297 (28.69)	924 (28.27)	373 (29.79)
History of chronic kidney disease	340 (7.52)	262 (8.01)	78 (6.23)
History of NMSC ^d	36 (0.80)	27 (0.83)	9 (0.72)
History of any cancer except NM-SC	3086 (68.26)	2189 (66.96)	897 (71.65)
Family history			
Family history of lung cancer	8 (0.18)	7 (0.21)	1 (0.08)
Family history of any cancer	79 (1.75)	63 (1.93)	16 (1.28)

^aGGO: ground-glass opacity.

^bEach patient can have more than 1 comorbidity.

^cCOPD: chronic obstructive pulmonary disease.

^dNMSC: nonmelanoma skin cancer.

Performance of the GGO NLP Pipeline

Among the cohort of 13,216 patients with lung cancer, 4521 (34.2%) had GGO reports, which comprised the “GGO cohort.” The NLP identified GGO features in 29,496 radiology notes of 4521 patients. Performance metrics for each GGO feature

are shown in [Table 2](#). The NLP pipeline achieved between 95% and 100% precision scores, 89% and 100% recall scores, and 92% and 100% F_1 -scores on different GGO features in the independent validation set. As an example, the GGO NLP algorithm correctly identified 986 pure GGOs out of 987 in the

gold standard and 145 part-solid GGOs out of 146 in the gold standard with a recall of 99.7% and 99%, respectively.

Table 2. Quality metrics of the NLP^a pipeline.

Semantic	Right ^b	Predict ^c	Gold ^d	Precision	Recall	F_1 -score
GGO ^e term: pure GGO	986	987	989	0.99	1	0.99
GGO term: part-solid GGO	145	146	146	0.99	0.99	0.99
GGO solidity	99	99	100	1	0.99	0.99
GGO shape/margin	144	151	144	0.95	1	0.98
GGO size	653	659	667	0.99	0.98	0.98
GGO quantity	154	156	160	0.99	0.96	0.97
GGO status change: better	46	46	46	1	1	1
GGO status change: worsen	107	107	110	1	0.97	0.99
GGO status change: stable	510	535	572	0.95	0.89	0.92
Potential GGO cause: infectious/inflammatory	146	147	148	0.99	0.99	0.99
Potential GGO cause: neoplasm	121	122	132	0.99	0.92	0.95
Potential GGO cause: others	71	73	76	0.97	0.95	0.95
GGO location	1164	1220	1270	0.95	0.92	0.93
Temporal	1650	1700	1650	0.97	1	0.99

^aNLP: natural language processing.

^bThe number of accurately extracted entities based on the gold standard.

^cThe number of entities predicted from the NLP pipeline.

^dManually annotated entity by annotators.

^eGGO: ground-glass opacity.

GGO Characteristics

Almost all patients (n=4432, 98%) had at least 1 pure GGO in their reports, and 11% (n=505) patients had terms related to part-solid GGOs. As shown in [Table 3](#), GGO location (3588/4521, 79.36%) was most often mentioned in notes and captured by NLP followed by potential etiology, GGO size, and change in GGO status. Over 60% (2277/3588, 63.46%) of patients had GGOs in both lungs, followed by the right lung only, with 43.42% (3948/9093 GGOs) of GGOs located in the upper lobes (Table S1 in [Multimedia Appendix 1](#)). Similarly, 43.80% (1095/2500) of patients had more than 1 potential etiology mentioned in their clinical notes, with the most common etiology being infectious or inflammatory. Around 10% (31/319)

of patients in the malignant neoplasm etiology group had precancerous conditions. Among the 2350 patients identified with data on GGO size, almost half of the patients had GGOs baseline size in the range category between 6 and 20 mm (1138/2350, 48.43%), followed by >20 mm (340/2350, 14.5%) and <6 mm (274/2350, 11.6%) categories. The vast majority (845/1043, 81%) of patients with reported GGO shape or margin indicated nodules with irregular or spiculated shape, and most patients seemed to have multiple GGOs (898/904, 99.3%) rather than single GGOs (6/904, 0.7%), but data for this attribute were not frequently captured in notes. The quantity entities, even when captured, were not described as integer values in most cases but as concept values such as numerous, scattered, and several.

Table 3. Distribution of NLP^a-identified GGO^b features in patients with GGO findings.

GGO attributes	Patients (N=4521), n (%)
Pure GGO	4432 (98)
Part solid GGO	505 (11)
Location^c	
Bilateral/both	2277 (63.5)
Left	438 (12.2)
Right	831 (23.2)
Unknown/subpleural	42 (1.1)
Potential etiology^c	
Infectious/inflammatory	795 (31.8)
Malignant neoplasm	319 (12.8)
Other	291 (11.6)
More than 1 cause	1095 (43.8)
Size^c	
<6 mm	274 (11.6)
6-20 mm	1139 (48.5)
>20 mm	340 (14.5)
More than 1 size	597 (25.4)
GGO status^c	
Better	97 (4.2)
Stable	1388 (59.4)
Worse	288 (12.3)
More than 1 status	564 (24.1)
Shape/margin^c	
Irregular/spiculated	845 (81)
Rounded/smooth	63 (6)
More than 1 shape	135 (13)
Change in GGO size^d	
Increase in size	240 (16.8)
Decrease in size	208 (14.6)
Stable in size	976 (68.5)
Change in GGO status^e	
Increased	259 (23)
Decreased	27 (2.4)
Stayed stable	815 (72.3)
Resolved	26 (2.3)

^aNLP: natural language processing.

^bGGO: ground-glass opacity.

^cPatient numbers were calculated from the first notes. GGO status was based on the description in the notes.

^dLongitudinal analysis between the first and the last notes.

^eLongitudinal analyses between the first and the subsequent notes.

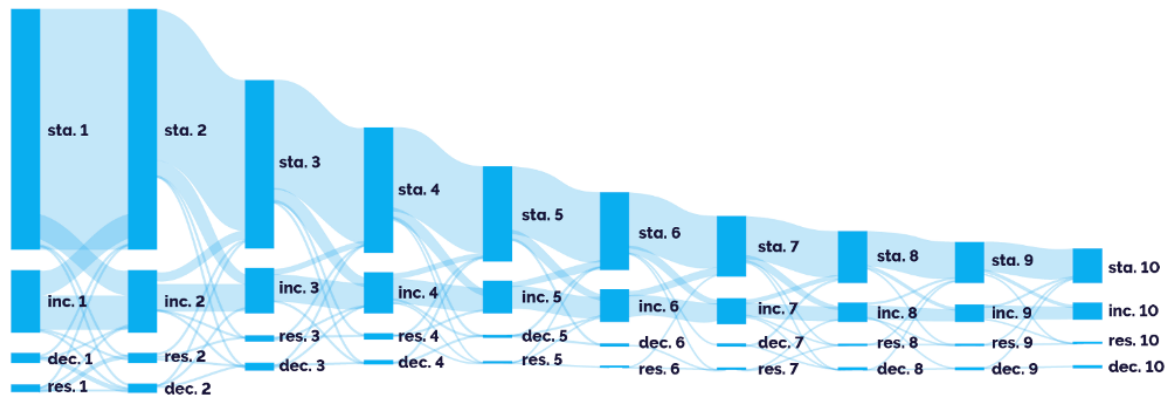
Longitudinal Analysis

Longitudinal analysis in patients with at least 2 GGO notes revealed that size increased in 16.8% (240/1424) of patients, decreased in 14.6% (208/1424), and remained unchanged in 68.5% (976/1424) in their last note compared to the first note (see Table 3 and Table S2 in Multimedia Appendix 1). The Figure S2 boxplot in Multimedia Appendix 1 shows GGO sizes at baseline and latest notes. Patients with GGO size available for only a single date were excluded from the plot. The largest GGO size was used if there was more than 1 size reported on the same day. The median GGO sizes among all relevant patients were smaller at the end point. We noticed that the patients starting with a large (>20 mm) baseline GGO size had a more medium/small GGO size reported at the end point compared with patients starting with a medium-sized GGO (see the bottom

right corner split by the red lines in Figure S2 in Multimedia Appendix 1).

A similar longitudinal analysis was performed to assess changes in GGOs over time, including indications in notes about changes in size and/or solidity or any descriptions of change. For this analysis, patients with more than 2 notes were included, and the most severe status change with the order of increased>stable>decrease was selected if more than 1 status change was reported in a day. Most patients (815/1127, 72.3%) had notes reporting a stable status of their GGOs, and “stable” was the only status reported for 40% (450/1127) of patients. The sequence of GGO status changes in the first 10 notes is depicted in Figure 4. For patients reported as stable, the subsequent report was usually stable again, followed by an increased status.

Figure 4. Analysis of ground-glass opacity (GGO) change in longitudinal notes. GGO status change (size and/or solidity) in the first 10 notes is visualized in the Sankey diagram. If a report had multiple status changes, the worst status change was selected. The majority of GGO stayed stable. Dec: decreased; Inc: increased; Res: resolved; Sta: stable.



Analysis of GGO Features and Interval Days Between GGO and Lung Cancer in the “Pregroup”

To examine whether our data are aligned with current knowledge about the impacts of size and location of nodules on lung malignancy, we analyzed GGOs in patients who had their first GGO reports before the lung cancer diagnosis date (called pregroup hereafter). Of 4521 patients with GGOs, 1706 (37.7%) were stratified into the pregroup. Among the 1706 pregroup patients, 853 (50%) patients had GGOs that can be classified exclusively into 1 baseline size group (<6 mm, 6-20 mm, or >20 mm). Table 4 shows the interval days between the first GGO report dates and the lung cancer diagnosis dates in each size group. We noted that 78% (136/174), 58% (319/550), and 47.3% (61/129) of patients had lung cancer diagnosis within 6

months in the >20 mm, 6 to 20 mm, and <6 mm groups, respectively. On the contrary, 16.6% (29/174), 31.5% (173/550), and 39.5% (51/129) of patients developed lung cancer after 1 year in the >20 mm, 6 to 20 mm, and <6 mm groups, respectively. Next, we investigated the location of GGOs in the pregroup. A total of 861 (50.5%) patients had a GGO location that could be classified into 1 location group (LLL, LUL, RLL, RML, or RUL). The upper lobe location was more frequently detected compared with the lower lobe location. Among the patients, 62.6% (539/861) had GGOs in the upper lobes, either RUL (336/861, 39%) or LUL (203/861, 23.6%). Moreover, 27.4% (236/861) of patients had GGOs in the lower lobes, either RLL (142/861, 16.5%) or LLL (94/861, 11%). The remaining 10% (86/861) of patients had GGOs in the middle lobe (RML).

Table 4. Patients in each size category at the different timelines from the first ground-glass opacity (GGO) notes to lung cancer diagnosis.

Size/timeline	<6 months, n (%)	6 months to 1 year, n (%)	1 year to 3 years, n (%)	>3 years, n (%)	Total, n (%)
<6 mm	61 (47.3)	17 (13.2)	29 (22.5)	22 (17)	129 (100)
6-20 mm	319 (58)	58 (10.5)	94 (17.1)	79 (14.4)	550 (100)
>20 mm	136 (78.2)	9 (5.2)	14 (8)	15 (8.6)	174 (100)

Discussion

Principal Findings

To understand the nature of GGOs in lung cancer cohorts, we constructed a GGO NLP pipeline in this study. Our data demonstrated high accuracy and efficiency of GGO feature identification for both pure GGOs and part-solid GGOs when this information was captured in notes. By implementing our model, we achieved automated extraction and analysis of GGO features in a huge volume of clinical notes, which enabled the identification of patients with GGOs for whom other clinical data were also available. Our model also enabled analysis of changes in GGO features over time by leveraging available longitudinal data at scale.

Similar to findings from Zheng et al [4] that utilized data from the community practices, we found that the laterality of the GGO nodules was more frequently documented in notes than other features like margins and shape. Hence, our study further supports the need for potentially standardizing the documentation of CT findings in radiology reports and progress notes. Early detection of GGOs and understanding of GGO features are critical for clinical decision-making, and they enable earlier intervention [51]. GGO status changes, including increased size and solidity, were described as critical factors for making a clinical decision on the resection [22]. Although a decrease in average nodule size has been observed across chest CT reports in general over time [4], in our study, we were able to use longitudinal data to track nodule changes specifically in each patient over time. Further analysis of whether this finding is related to treating larger GGOs can provide a better interpretation of this result and insights into GGO treatment. In our study, we also observed that the majority of patients with a GGO larger than 20 mm were diagnosed with lung cancer in the 6 months following the GGO finding.

Although GGO solidity information is one of the most critical prognostic factors [52], except for the pure or part-solid GGO information, additional GGO solidity information—such as absolute solid component sizes or solidity status changes—was not automatically extracted in previous NLP studies. In this study, we showed the feasibility of tracking the solidity status changes, as captured in the notes, but changes in every nodule may not be reflected. The solidity status changes including density change were curated by comparing the baseline and last note GGO terms. Our data revealed that most patients with solidity change information showed either a solidity increase (from pure to part solid) or stayed stable.

The quantity of GGO nodules is another crucial piece of information. It has been found that 1 to 4 GGO in a single note can be cancerous with no significant difference between 1 to 4 nodules, but ≥ 5 is more likely infectious/inflammatory in the etiology [53,54]. In many notes, the entities indicating the total

number of GGO were not found. Radiologists described the number of GGO nodules as concepts like numerous or scattered rather than giving the actual number of GGO nodules when there are multiple GGO. Although we classified the number of GGO as multiple or single in this study, further subtyping the number of GGO nodules as 1 to 4 or ≥ 5 in future work by counting each GGO term extracted and their related attributes, such as location and size, could provide better insights.

Strengths and Limitations

Although NLP technologies have significantly impacted real-world evidence generation, there remain unmet needs in clinical data retrieval such as relation recognition, longitudinal analysis, and providing insights rather than extracting data only, as Sheikhalishahi et al [39] described in their systematic review. In our deep learning model, we showed the feasibility of relation extraction rather than isolated entity extraction only and the temporal reasoning for the longitudinal analysis of patient-level data analysis. Transformer models such as BERT-based models can be examined together in future work.

There are several limitations to our study. We analyzed the GGO data in a lung cancer cohort for the initial feasibility assessment. However, our NLP pipeline can be easily expanded to other cohorts such as non-lung cancer cohorts with GGO reports in future studies, which provides more opportunities such as analyzing the associated risk factors of developing lung cancer from GGO. Additionally, a deeper analysis of pre- and postdiagnosis patient journeys can provide more insights into preventing and detecting lung malignancy. In radiology reports with multiple GGOs, tracking individual GGOs across reports over time for the longitudinal analysis of individual GGOs is challenging. Further efforts for identifying and monitoring each GGO can give us better insights into each GGO's nature and outcome. NLP is naturally limited by its ability to capture only documented information. However, Zheng et al [4] reported trends of increasing documentation of smaller nodules and their features in radiology reports. Given this fact, NLP can be utilized as a powerful tool to study the natural history of GGOs and identify cohorts of interest for further analysis or for more in-depth radiomics work.

Conclusions

Our study demonstrates that the deep NLP model can automatically extract granular GGO features, when documented, at scale. The model could be deployed further to large volumes of longitudinal free-text reports to continuously update prognosis as an individual's disease course unfolds and leverage the longitudinal data with treatment patterns, clinical outcomes, and risk factors for various applications. The AI-enabled model offers a potential advantage as an automated clinical decision support tool to identify cohorts of interest for radiomics and optimize resource utilization for cancer prevention, early detection, and effective management.

Acknowledgments

We thank Arielle Redfern, Jeremie Carlson, Emily Reed, Rene Dempsey, and Hui Kim for helping us with the annotation. We also thank Tony Prentice, Tom Neyarapally, Anatol Blass, Aaron Black, Paul McDonagh, Aaron Zhang, and the data curation team who made this study possible.

Data Availability

The data used in this study are not open access due to patient privacy, security, and Health Insurance Portability and Accountability Act (HIPAA) requirements. To enable a complete run of the code shared in this study, a minimum amount of desensitized sample data could be shared with the sharing agreement. Relevant requests should be addressed to author ZL (zongzhi.liu@sema4.com). The source code of this study is provided on the GitHub website under the search term “ground glass opacity (GGO).”

Authors' Contributions

KL, ZL, UC, IK, and XW designed the study and wrote the manuscript. KL and XW reviewed the literature and patient notes and constructed the ground-glass opacity ontology. KL, ZL, MM, ML, YM, CG, TW, UC, and BL were involved in the model training, postprocessing, and data analysis. MKH, TJ, BL, LA, PA, QP, WO, GS, ES, and XW discussed the project and reviewed the manuscript.

Conflicts of Interest

KL, ZL, TJ, MM, ML, YM, CG, TW, LA, PA, QP, and XW are employees of Sema4. UC, IK, and BL are employees of Johnson & Johnson. WO and ES are employees of the Icahn School of Medicine at Mount Sinai. WO receives equity from Sema4 and GeneDx. MKH is an employee of GeneDx and receives equity as part of compensation. All authors declare no other competing financial or nonfinancial interests.

Multimedia Appendix 1

Additional figures and tables showing duration distribution between ground-glass opacity (GGO) reports and lung cancer, analytics output of GGO size change, GGO location distribution, and longitudinal analysis of GGO size changes.

[\[DOCX File , 176 KB - ai_v2i1e44537_app1.docx \]](#)

References

1. Shieh Y, Eklund M, Sawaya GF, Black WC, Kramer BS, Esserman LJ. Population-based screening for cancer: hope and hype. *Nat Rev Clin Oncol* 2016 Sep;13(9):550-565 [FREE Full text] [doi: [10.1038/nrclinonc.2016.50](https://doi.org/10.1038/nrclinonc.2016.50)] [Medline: [27071351](https://pubmed.ncbi.nlm.nih.gov/27071351/)]
2. Field JK, Vulkan D, Davies MP, Baldwin DR, Brain KE, Devaraj A, et al. Lung cancer mortality reduction by LDCT screening: UKLS randomised trial results and international meta-analysis. *Lancet Reg Health Eur* 2021 Nov;10:100179 [FREE Full text] [doi: [10.1016/j.lanepe.2021.100179](https://doi.org/10.1016/j.lanepe.2021.100179)] [Medline: [34806061](https://pubmed.ncbi.nlm.nih.gov/34806061/)]
3. National Lung Screening Trial Research Team, Aberle DR, Adams AM, Berg CD, Black WC, Clapp JD, et al. Reduced lung-cancer mortality with low-dose computed tomographic screening. *N Engl J Med* 2011 Aug 04;365(5):395-409 [FREE Full text] [doi: [10.1056/NEJMoa1102873](https://doi.org/10.1056/NEJMoa1102873)] [Medline: [21714641](https://pubmed.ncbi.nlm.nih.gov/21714641/)]
4. Zheng C, Huang BZ, Agazaryan AA, Creekmur B, Osuj TA, Gould MK. Natural Language Processing to Identify Pulmonary Nodules and Extract Nodule Characteristics From Radiology Reports. *Chest* 2021 Nov;160(5):1902-1914. [doi: [10.1016/j.chest.2021.05.048](https://doi.org/10.1016/j.chest.2021.05.048)] [Medline: [34089738](https://pubmed.ncbi.nlm.nih.gov/34089738/)]
5. Gould MK, Tang T, Liu IA, Lee J, Zheng C, Danforth KN, et al. Recent trends in the identification of incidental pulmonary nodules. *Am J Respir Crit Care Med* 2015 Nov 15;192(10):1208-1214. [doi: [10.1164/rccm.201505-0990OC](https://doi.org/10.1164/rccm.201505-0990OC)] [Medline: [26214244](https://pubmed.ncbi.nlm.nih.gov/26214244/)]
6. Mazzone PJ, Lam L. Evaluating the patient with a pulmonary nodule: a review. *JAMA* 2022 Jan 18;327(3):264-273. [doi: [10.1001/jama.2021.24287](https://doi.org/10.1001/jama.2021.24287)] [Medline: [35040882](https://pubmed.ncbi.nlm.nih.gov/35040882/)]
7. Pedersen JH, Saghir Z, Wille MMW, Thomsen LH, Skov BG, Ashraf H. Ground-glass opacity lung nodules in the era of lung cancer CT screening: radiology, pathology, and clinical management. *Oncology (Williston Park)* 2016 Mar;30(3):266-274 [FREE Full text] [Medline: [26984222](https://pubmed.ncbi.nlm.nih.gov/26984222/)]
8. Chen K, Bai J, Reuben A, Zhao H, Kang G, Zhang C, et al. Multiomics analysis reveals distinct immunogenomic features of lung cancer with ground-glass opacity. *Am J Respir Crit Care Med* 2021 Nov 15;204(10):1180-1192 [FREE Full text] [doi: [10.1164/rccm.202101-0119OC](https://doi.org/10.1164/rccm.202101-0119OC)] [Medline: [34473939](https://pubmed.ncbi.nlm.nih.gov/34473939/)]
9. Kim YW, Lee C. Optimal management of pulmonary ground-glass opacity nodules. *Transl Lung Cancer Res* 2019 Dec;8(Suppl 4):S418-S424 [FREE Full text] [doi: [10.21037/tlcr.2019.10.24](https://doi.org/10.21037/tlcr.2019.10.24)] [Medline: [32038928](https://pubmed.ncbi.nlm.nih.gov/32038928/)]
10. Kobayashi Y, Mitsudomi T. Management of ground-glass opacities: should all pulmonary lesions with ground-glass opacity be surgically resected? *Transl Lung Cancer Res* 2013 Oct;2(5):354-363 [FREE Full text] [doi: [10.3978/j.issn.2218-6751.2013.09.03](https://doi.org/10.3978/j.issn.2218-6751.2013.09.03)] [Medline: [25806254](https://pubmed.ncbi.nlm.nih.gov/25806254/)]

11. Zhang Y, Fu F, Chen H. Management of ground-glass opacities in the lung cancer spectrum. *Ann Thorac Surg* 2020 Dec;110(6):1796-1804. [doi: [10.1016/j.athoracsur.2020.04.094](https://doi.org/10.1016/j.athoracsur.2020.04.094)] [Medline: [32525031](https://pubmed.ncbi.nlm.nih.gov/32525031/)]
12. Jang HJ, Lee KS, Kwon OJ, Rhee CH, Shim YM, Han J. Bronchioloalveolar carcinoma: focal area of ground-glass attenuation at thin-section CT as an early sign. *Radiology* 1996 May;199(2):485-488. [doi: [10.1148/radiology.199.2.8668800](https://doi.org/10.1148/radiology.199.2.8668800)] [Medline: [8668800](https://pubmed.ncbi.nlm.nih.gov/8668800/)]
13. Migliore M, Fornito M, Palazzolo M, Criscione A, Gangemi M, Borrata F, et al. Ground glass opacities management in the lung cancer screening era. *Ann Transl Med* 2018 Mar;6(5):90 [FREE Full text] [doi: [10.21037/atm.2017.07.28](https://doi.org/10.21037/atm.2017.07.28)] [Medline: [29666813](https://pubmed.ncbi.nlm.nih.gov/29666813/)]
14. Van Haren RM, Correa AM, Sepesi B, Rice DC, Hofstetter WL, Mehran RJ, et al. Ground glass lesions on chest imaging: evaluation of reported incidence in cancer patients using natural language processing. *Ann Thorac Surg* 2019 Mar;107(3):936-940. [doi: [10.1016/j.athoracsur.2018.09.016](https://doi.org/10.1016/j.athoracsur.2018.09.016)] [Medline: [30612991](https://pubmed.ncbi.nlm.nih.gov/30612991/)]
15. Bach PB, Mirkin JN, Oliver TK, Azzoli CG, Berry DA, Brawley OW, et al. Benefits and harms of CT screening for lung cancer: a systematic review. *JAMA* 2012 Jun 13;307(22):2418-2429 [FREE Full text] [doi: [10.1001/jama.2012.5521](https://doi.org/10.1001/jama.2012.5521)] [Medline: [22610500](https://pubmed.ncbi.nlm.nih.gov/22610500/)]
16. Goo JM, Park CM, Lee HJ. Ground-glass nodules on chest CT as imaging biomarkers in the management of lung adenocarcinoma. *Am J Roentgenol* 2011 Mar;196(3):533-543. [doi: [10.2214/ajr.10.5813](https://doi.org/10.2214/ajr.10.5813)]
17. Infante M, Lutman RF, Imparato S, Di Rocco M, Ceresoli GL, Torri V, et al. Differential diagnosis and management of focal ground-glass opacities. *Eur Respir J* 2009 Apr 01;33(4):821-827 [FREE Full text] [doi: [10.1183/09031936.00047908](https://doi.org/10.1183/09031936.00047908)] [Medline: [19047318](https://pubmed.ncbi.nlm.nih.gov/19047318/)]
18. Kim HY, Shim YM, Lee KS, Han J, Yi CA, Kim YK. Persistent pulmonary nodular ground-glass opacity at thin-section CT: histopathologic comparisons. *Radiology* 2007 Oct;245(1):267-275. [doi: [10.1148/radiol.2451061682](https://doi.org/10.1148/radiol.2451061682)] [Medline: [17885195](https://pubmed.ncbi.nlm.nih.gov/17885195/)]
19. Naidich DP, Bankier AA, MacMahon H, Schaefer-Prokop CM, Pistolesi M, Goo JM, et al. Recommendations for the management of subsolid pulmonary nodules detected at CT: a statement from the Fleischner Society. *Radiology* 2013 Jan;266(1):304-317. [doi: [10.1148/radiol.12120628](https://doi.org/10.1148/radiol.12120628)] [Medline: [23070270](https://pubmed.ncbi.nlm.nih.gov/23070270/)]
20. Henschke CI, Yankelevitz DF, Mirtcheva R, McGuinness G, McCauley D, Miettinen OS, ELCAP Group. CT screening for lung cancer: frequency and significance of part-solid and nonsolid nodules. *AJR Am J Roentgenol* 2002 May;178(5):1053-1057. [doi: [10.2214/ajr.178.5.1781053](https://doi.org/10.2214/ajr.178.5.1781053)] [Medline: [11959700](https://pubmed.ncbi.nlm.nih.gov/11959700/)]
21. Khan T, Usman Y, Abdo T, Chaudry F, Keddissi JL, Youness HA. Diagnosis and management of peripheral lung nodule. *Ann Transl Med* 2019 Aug;7(15):348-348 [FREE Full text] [doi: [10.21037/atm.2019.03.59](https://doi.org/10.21037/atm.2019.03.59)] [Medline: [31516894](https://pubmed.ncbi.nlm.nih.gov/31516894/)]
22. MacMahon H, Naidich DP, Goo JM, Lee KS, Leung ANC, Mayo JR, et al. Guidelines for management of incidental pulmonary nodules detected on CT images: from the Fleischner Society 2017. *Radiology* 2017 Jul;284(1):228-243. [doi: [10.1148/radiol.2017161659](https://doi.org/10.1148/radiol.2017161659)] [Medline: [28240562](https://pubmed.ncbi.nlm.nih.gov/28240562/)]
23. McWilliams A, Tammemagi MC, Mayo JR, Roberts H, Liu G, Soghrati K, et al. Probability of cancer in pulmonary nodules detected on first screening CT. *N Engl J Med* 2013 Sep 05;369(10):910-919 [FREE Full text] [doi: [10.1056/NEJMoa1214726](https://doi.org/10.1056/NEJMoa1214726)] [Medline: [24004118](https://pubmed.ncbi.nlm.nih.gov/24004118/)]
24. Loverdos K, Fotiadis A, Kontogianni C, Iliopoulou M, Gaga M. Lung nodules: A comprehensive review on current approach and management. *Ann Thorac Med* 2019;14(4):226-238 [FREE Full text] [doi: [10.4103/atm.ATM_110_19](https://doi.org/10.4103/atm.ATM_110_19)] [Medline: [31620206](https://pubmed.ncbi.nlm.nih.gov/31620206/)]
25. Hu D, Li S, Zhang H, Wu N, Lu X. Using natural language processing and machine learning to preoperatively predict lymph node metastasis for non-small cell lung cancer with electronic medical records: development and validation study. *JMIR Med Inform* 2022 Apr 25;10(4):e35475 [FREE Full text] [doi: [10.2196/35475](https://doi.org/10.2196/35475)] [Medline: [35468085](https://pubmed.ncbi.nlm.nih.gov/35468085/)]
26. Li W, Wang X, Zhang Y, Li X, Li Q, Ye Z. Radiomic analysis of pulmonary ground-glass opacity nodules for distinction of preinvasive lesions, invasive pulmonary adenocarcinoma and minimally invasive adenocarcinoma based on quantitative texture analysis of CT. *Chin J Cancer Res* 2018 Aug;30(4):415-424 [FREE Full text] [doi: [10.21147/j.issn.1000-9604.2018.04.04](https://doi.org/10.21147/j.issn.1000-9604.2018.04.04)] [Medline: [30210221](https://pubmed.ncbi.nlm.nih.gov/30210221/)]
27. van Timmeren JE, Cester D, Tanadini-Lang S, Alkadhi H, Baessler B. Radiomics in medical imaging-"how-to" guide and critical reflection. *Insights Imaging* 2020 Aug 12;11(1):91 [FREE Full text] [doi: [10.1186/s13244-020-00887-2](https://doi.org/10.1186/s13244-020-00887-2)] [Medline: [32785796](https://pubmed.ncbi.nlm.nih.gov/32785796/)]
28. Ibrahim A, Primakov S, Beuque M, Woodruff H, Halilaj I, Wu G, et al. Radiomics for precision medicine: current challenges, future prospects, and the proposal of a new framework. *Methods* 2021 Apr;188:20-29 [FREE Full text] [doi: [10.1016/j.ymeth.2020.05.022](https://doi.org/10.1016/j.ymeth.2020.05.022)] [Medline: [32504782](https://pubmed.ncbi.nlm.nih.gov/32504782/)]
29. Delli Pizzi A, Chiarelli AM, Chiacchiaretta P, Valdesi C, Croce P, Mastrodicasa D, et al. Radiomics-based machine learning differentiates "ground-glass" opacities due to COVID-19 from acute non-COVID-19 lung disease. *Sci Rep* 2021 Aug 26;11(1):17237 [FREE Full text] [doi: [10.1038/s41598-021-96755-0](https://doi.org/10.1038/s41598-021-96755-0)] [Medline: [34446812](https://pubmed.ncbi.nlm.nih.gov/34446812/)]
30. Liu Z, Wang S, Dong D, Wei J, Fang C, Zhou X, et al. The applications of radiomics in precision diagnosis and treatment of oncology: opportunities and challenges. *Theranostics* 2019;9(5):1303-1322 [FREE Full text] [doi: [10.7150/thno.30309](https://doi.org/10.7150/thno.30309)] [Medline: [30867832](https://pubmed.ncbi.nlm.nih.gov/30867832/)]

31. Rizzo S, Botta F, Raimondi S, Origgi D, Fanciullo C, Morganti AG, et al. Radiomics: the facts and the challenges of image analysis. *Eur Radiol Exp* 2018 Nov 14;2(1):36 [FREE Full text] [doi: [10.1186/s41747-018-0068-z](https://doi.org/10.1186/s41747-018-0068-z)] [Medline: [30426318](https://pubmed.ncbi.nlm.nih.gov/30426318/)]
32. Limkin E, Sun R, Dercle L, Zacharaki E, Robert C, Reuzé S, et al. Promises and challenges for the implementation of computational medical imaging (radiomics) in oncology. *Ann Oncol* 2017 Jun 01;28(6):1191-1206 [FREE Full text] [doi: [10.1093/annonc/mdx034](https://doi.org/10.1093/annonc/mdx034)] [Medline: [28168275](https://pubmed.ncbi.nlm.nih.gov/28168275/)]
33. Huang S, Pareek A, Seyyedi S, Banerjee I, Lungren MP. Fusion of medical imaging and electronic health records using deep learning: a systematic review and implementation guidelines. *NPJ Digit Med* 2020;3:136 [FREE Full text] [doi: [10.1038/s41746-020-00341-z](https://doi.org/10.1038/s41746-020-00341-z)] [Medline: [33083571](https://pubmed.ncbi.nlm.nih.gov/33083571/)]
34. Pons E, Braun LMM, Hunink MGM, Kors JA. Natural language processing in radiology: a systematic review. *Radiology* 2016 May;279(2):329-343. [doi: [10.1148/radiol.16142770](https://doi.org/10.1148/radiol.16142770)] [Medline: [27089187](https://pubmed.ncbi.nlm.nih.gov/27089187/)]
35. Beyer SE, McKee BJ, Regis SM, McKee AB, Flacke S, El Saadawi G, et al. Automatic Lung-RADS classification with a natural language processing system. *J Thorac Dis* 2017 Sep;9(9):3114-3122 [FREE Full text] [doi: [10.21037/jtd.2017.08.13](https://doi.org/10.21037/jtd.2017.08.13)] [Medline: [29221286](https://pubmed.ncbi.nlm.nih.gov/29221286/)]
36. Cook MJ, Yao L, Wang X. Facilitating accurate health provider directories using natural language processing. *BMC Med Inform Decis Mak* 2019 Apr 04;19(Suppl 3):80 [FREE Full text] [doi: [10.1186/s12911-019-0788-x](https://doi.org/10.1186/s12911-019-0788-x)] [Medline: [30943977](https://pubmed.ncbi.nlm.nih.gov/30943977/)]
37. Dave AD, Ruano G, Kost J, Wang X. Automated extraction of pain symptoms: a natural language approach using electronic health records. *Pain Physician* 2022 Mar;25(2):E245-E254 [FREE Full text] [Medline: [35322976](https://pubmed.ncbi.nlm.nih.gov/35322976/)]
38. Wang X, Hripcsak G, Markatou M, Friedman C. Active computerized pharmacovigilance using natural language processing, statistics, and electronic health records: a feasibility study. *JAMA* 2009 May 01;16(3):328-337. [doi: [10.1197/jamia.m3028](https://doi.org/10.1197/jamia.m3028)]
39. Sheikhalishahi S, Miotto R, Dudley JT, Lavelli A, Rinaldi F, Osmani V. Natural language processing of clinical notes on chronic diseases: systematic review. *JMIR Med Inform* 2019 Apr 27;7(2):e12239 [FREE Full text] [doi: [10.2196/12239](https://doi.org/10.2196/12239)] [Medline: [31066697](https://pubmed.ncbi.nlm.nih.gov/31066697/)]
40. Tsuji S, Wen A, Takahashi N, Zhang H, Ogasawara K, Jiang G. Developing a RadLex-based named entity recognition tool for mining textual radiology reports: development and performance evaluation study. *J Med Internet Res* 2021 Oct 29;23(10):e25378 [FREE Full text] [doi: [10.2196/25378](https://doi.org/10.2196/25378)] [Medline: [34714247](https://pubmed.ncbi.nlm.nih.gov/34714247/)]
41. Linna N, Kahn CE. Applications of natural language processing in radiology: A systematic review. *Int J Med Inform* 2022 Jul;163:104779. [doi: [10.1016/j.ijmedinf.2022.104779](https://doi.org/10.1016/j.ijmedinf.2022.104779)] [Medline: [35533413](https://pubmed.ncbi.nlm.nih.gov/35533413/)]
42. Liu H, Zhang Z, Xu Y, Wang N, Huang Y, Yang Z, et al. Use of BERT (Bidirectional Encoder Representations from Transformers)-based deep learning method for extracting evidences in Chinese radiology reports: development of a computer-aided liver cancer diagnosis framework. *J Med Internet Res* 2021 Jan 12;23(1):e19689 [FREE Full text] [doi: [10.2196/19689](https://doi.org/10.2196/19689)] [Medline: [33433395](https://pubmed.ncbi.nlm.nih.gov/33433395/)]
43. Farjah F, Halgrim S, Buist DSM, Gould MK, Zeliadt SB, Loggers ET, et al. An automated method for identifying individuals with a lung nodule can be feasibly implemented across health systems. *EGEMS (Wash DC)* 2016;4(1):1254 [FREE Full text] [doi: [10.13063/2327-9214.1254](https://doi.org/10.13063/2327-9214.1254)] [Medline: [27668266](https://pubmed.ncbi.nlm.nih.gov/27668266/)]
44. Kang SK, Garry K, Chung R, Moore WH, Iturrate E, Swartz JL, et al. Natural language processing for identification of incidental pulmonary nodules in radiology reports. *J Am Coll Radiol* 2019 Nov;16(11):1587-1594 [FREE Full text] [doi: [10.1016/j.jacr.2019.04.026](https://doi.org/10.1016/j.jacr.2019.04.026)] [Medline: [31132331](https://pubmed.ncbi.nlm.nih.gov/31132331/)]
45. Shewale JB, Nelson DB, Rice DC, Sepesi B, Hofstetter WL, Mehran RJ, et al. Natural history of ground-glass lesions among patients with previous lung cancer. *Ann Thorac Surg* 2018 Jun;105(6):1671-1677. [doi: [10.1016/j.athoracsur.2018.01.031](https://doi.org/10.1016/j.athoracsur.2018.01.031)] [Medline: [29432718](https://pubmed.ncbi.nlm.nih.gov/29432718/)]
46. Danforth KN, Early MI, Ngan S, Kosco AE, Zheng C, Gould MK. Automated identification of patients with pulmonary nodules in an integrated health system using administrative health plan data, radiology reports, and natural language processing. *J Thorac Oncol* 2012 Aug;7(8):1257-1262 [FREE Full text] [doi: [10.1097/JTO.0b013e31825bd9f5](https://doi.org/10.1097/JTO.0b013e31825bd9f5)] [Medline: [22627647](https://pubmed.ncbi.nlm.nih.gov/22627647/)]
47. Mikolov T, Chen K, Corrado G, Dean J. Efficient Estimation of Word Representations in Vector Space. arXiv Preprint posted online January 16, 2013. [doi: [10.48550/arXiv.1301.3781](https://doi.org/10.48550/arXiv.1301.3781)]
48. Soysal E, Wang J, Jiang M, Wu Y, Pakhomov S, Liu H, et al. CLAMP - a toolkit for efficiently building customized clinical natural language processing pipelines. *J Am Med Inform Assoc* 2018 Mar 01;25(3):331-336 [FREE Full text] [doi: [10.1093/jamia/ocx132](https://doi.org/10.1093/jamia/ocx132)] [Medline: [29186491](https://pubmed.ncbi.nlm.nih.gov/29186491/)]
49. Alfattni G, Peek N, Nenadic G. Attention-based bidirectional long short-term memory networks for extracting temporal relationships from clinical discharge summaries. *J Biomed Inform* 2021 Nov;123:103915 [FREE Full text] [doi: [10.1016/j.jbi.2021.103915](https://doi.org/10.1016/j.jbi.2021.103915)] [Medline: [34600144](https://pubmed.ncbi.nlm.nih.gov/34600144/)]
50. Xu J, Li Z, Wei Q, Wu Y, Xiang Y, Lee H, et al. Applying a deep learning-based sequence labeling approach to detect attributes of medical concepts in clinical text. *BMC Med Inform Decis Mak* 2019 Dec 05;19(Suppl 5):236 [FREE Full text] [doi: [10.1186/s12911-019-0937-2](https://doi.org/10.1186/s12911-019-0937-2)] [Medline: [31801529](https://pubmed.ncbi.nlm.nih.gov/31801529/)]
51. Ost DE, Gould MK. Decision making in patients with pulmonary nodules. *Am J Respir Crit Care Med* 2012 Feb 15;185(4):363-372 [FREE Full text] [doi: [10.1164/rccm.201104-0679CI](https://doi.org/10.1164/rccm.201104-0679CI)] [Medline: [21980032](https://pubmed.ncbi.nlm.nih.gov/21980032/)]

52. Kakinuma R, Noguchi M, Ashizawa K, Kuriyama K, Maeshima AM, Koizumi N, et al. Natural history of pulmonary subsolid nodules: a prospective multicenter study. *J Thorac Oncol* 2016 Jul;11(7):1012-1028 [[FREE Full text](#)] [doi: [10.1016/j.jtho.2016.04.006](https://doi.org/10.1016/j.jtho.2016.04.006)] [Medline: [27089851](#)]
53. Peters R, Heuvelmans MA, van Ooijen PM, De Bock GH, Oudkerk M. Prevalence of pulmonary multi-nodularity in CT lung cancer screening and lung cancer probability. 2015 Presented at: Radiological Society of North America Scientific Assembly and Annual Meeting 2015; November 29-December 4; Oak Brook, IL p. 111-111. [doi: [10.1102/1470-7330.2013.9043](https://doi.org/10.1102/1470-7330.2013.9043)]
54. Heuvelmans MA, Walter JE, Peters RB, Bock GHD, Yousaf-Khan U, Aalst CMVD, et al. Relationship between nodule count and lung cancer probability in baseline CT lung cancer screening: The NELSON study. *Lung Cancer* 2017 Nov;113:45-50 [[FREE Full text](#)] [doi: [10.1016/j.lungcan.2017.08.023](https://doi.org/10.1016/j.lungcan.2017.08.023)] [Medline: [29110848](#)]

Abbreviations

AI: artificial intelligence
BERT: Bidirectional Encoder Representations From Transformers
Bi-LSTM: bidirectional long-short term memory
CLAMP: Clinical Language Annotation, Modeling, And Processing
CRF: conditional random field
CT: computed tomography
EHR: electronic health record
FN: false negative
FP: false positive
GGO: ground-glass opacity
ICD: International Classification of Diseases
LLL: left lower lobe
LUL: left upper lobe
NLP: natural language processing
PPV: positive predictive value
RLL: right lower lobe
RML: right middle lobe
RUL: right upper lobe
TP: true positive

Edited by H Liu; submitted 23.11.22; peer-reviewed by H Wang, K Gupta, N Jiwani, M Elbattah; comments to author 26.12.22; revised version received 30.01.23; accepted 31.03.23; published 01.06.23.

Please cite as:

Lee K, Liu Z, Chandran U, Kalsekar I, Laxmanan B, Higashi MK, Jun T, Ma M, Li M, Mai Y, Gilman C, Wang T, Ai L, Aggarwal P, Pan Q, Oh W, Stolovitzky G, Schadt E, Wang X

Detecting Ground Glass Opacity Features in Patients With Lung Cancer: Automated Extraction and Longitudinal Analysis via Deep Learning-Based Natural Language Processing

JMIR AI 2023;2:e44537

URL: <https://ai.jmir.org/2023/1/e44537>

doi: [10.2196/44537](https://doi.org/10.2196/44537)

PMID:

©Kyeryoung Lee, Zongzhi Liu, Urmila Chandran, Iftexhar Kalsekar, Balaji Laxmanan, Mitchell K Higashi, Tomi Jun, Meng Ma, Minghao Li, Yun Mai, Christopher Gilman, Tongyu Wang, Lei Ai, Parag Aggarwal, Qi Pan, William Oh, Gustavo Stolovitzky, Eric Schadt, Xiaoyan Wang. Originally published in JMIR AI (<https://ai.jmir.org>), 01.06.2023. This is an open-access article distributed under the terms of the Creative Commons Attribution License (<https://creativecommons.org/licenses/by/4.0/>), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR AI, is properly cited. The complete bibliographic information, a link to the original publication on <https://www.ai.jmir.org/>, as well as this copyright and license information must be included.

Original Paper

A Trainable Open-Source Machine Learning Accelerometer Activity Recognition Toolbox: Deep Learning Approach

Fluri Wieland¹, BSc, MSc; Claudio Nigg¹, BSc, MSc, PhD

Department of Health Science, Institute of Sports Science, University of Bern, Bern, Switzerland

Corresponding Author:

Fluri Wieland, BSc, MSc
Department of Health Science
Institute of Sports Science
University of Bern
Bremgartenstrasse 145
Bern, 3012
Switzerland
Phone: 41 787347220
Email: flu.wieland@gmail.com

Abstract

Background: The accuracy of movement determination software in current activity trackers is insufficient for scientific applications, which are also not open-source.

Objective: To address this issue, we developed an accurate, trainable, and open-source smartphone-based activity-tracking toolbox that consists of an Android app (*HumanActivityRecorder*) and 2 different deep learning algorithms that can be adapted to new behaviors.

Methods: We employed a semisupervised deep learning approach to identify the different classes of activity based on accelerometry and gyroscope data, using both our own data and open competition data.

Results: Our approach is robust against variation in sampling rate and sensor dimensional input and achieved an accuracy of around 87% in classifying 6 different behaviors on both our own recorded data and the MotionSense data. However, if the dimension-adaptive neural architecture model is tested on our own data, the accuracy drops to 26%, which demonstrates the superiority of our algorithm, which performs at 63% on the MotionSense data used to train the dimension-adaptive neural architecture model.

Conclusions: *HumanActivityRecorder* is a versatile, retrainable, open-source, and accurate toolbox that is continually tested on new data. This enables researchers to adapt to the behavior being measured and achieve repeatability in scientific studies.

(JMIR AI 2023;2:e42337) doi:[10.2196/42337](https://doi.org/10.2196/42337)

KEYWORDS

activity classification; deep learning; accelerometry; open source; activity recognition; machine learning; activity recorder; digital health application; smartphone app; deep learning algorithm; sensor device

Introduction

Background

The last decade has seen a significant increase in worldwide smartphone ownership [1], with approximately half of the world's population now owning a smartphone and a device penetration rate of 80% in Germany and the United Kingdom [2]. Even low-end smartphones are equipped with various sensors, including accelerometers, gyroscopes, proximity sensors, magnetometers, and GPS receivers, along with energy-efficient processors and stable internet connections.

With the advent of smartphones and wearables, physical activity analysis has greatly gained in popularity. Accelerometry-based behavior analysis has a variety of applications, such as fall detection in older patients [3], health monitoring [4], work-related stress analysis [5], and sleep analysis [6]. The widespread use of accelerometry in everyday smartphone apps has reduced the cost of gyroscope and accelerometer sensors, which has in turn accelerated their development. While wearables have gained popularity as accelerometer devices, smartphones still make up the majority of them.

Many studies have shown the accuracy and reliability of smartphone sensors in accelerometry [7-9]. Although wearables tend to provide more accurate behavior classifications, the potential of using smartphones far outweighs the additional accuracy gained from wearables. Although they are more precise thus far [10], the cost of wearables for larger study populations is very high, compared with the widespread popularity and affordability of smartphones, making them a more accessible option for research. Additionally, smartphone apps are easier to distribute, update, configure, and adapt to specific research questions than wearables. Wearables also have the disadvantage of limited software support and closed-source software, making research based on previous software nonreproducible after algorithm updates. This means that wearables bought for research purposes must be replaced on a regular basis.

Most importantly, however, the default software of wearable manufacturers is in almost all cases not open-source, meaning that after each change of the algorithm (ie, app update) that classifies behavior, research based on previous software is not reproducible anymore. Furthermore, in most cases, charges apply for the use of the said software. On the other hand, some smartphone manufacturers offer free, open-source toolboxes for movement activity recognition, such as Samsung and Huawei. However, these toolboxes only recognize a limited number of activity types and are at the time of writing not trainable to new activities. The purpose of both, however, is for them to be integrated into applications, so they can be used to determine whether a smartphone user is moving and is active or not, in order to interact with application functionality, such as energy saving while not moving, clocking active hours, or encouraging movement when a user is inactive. While data can be collected and stored, the behavior classes are fixed and neither trainable nor retrainable. To address these limitations, the scientific community needs access to an open-source, adaptable behavior analysis toolbox that also facilitates reproducible research and is adaptable to specific research questions. To fulfil this need, we present our open-source, deep learning-based behavior analysis toolbox. Our Human Activity Analysis toolbox includes a proprietary Android app, 2 deep learning algorithms, scripts to process data, and a continually expanding sample data set. The toolbox has been validated with a sample of 68 University of Bern students and employees.

Activity Recognition and Deep Learning Background

Deep learning algorithms have gained importance in classifying human behavior based on sensor data collected from accelerometers, gyroscopes, and magnetometers [11-18] (for a deeper understanding and comprehensive overview, see [19]). These algorithms are based on artificial neural networks, and specifically, deep neural networks (DNNs) have become the dominant approach for activity recognition as of 2022. DNNs consist of multiple layers of neurons of similar or different types, and the functionality of these neurons is determined by the nature of the layers and the way they are interconnected [20,21]. It is important to note that a standard neural network consists of many simple, connected processors called neurons, each producing a sequence of real-valued activations. Depending on the problem and how the neurons are connected, such behavior may require long causal chains of computational stages. Thus,

if multiple layers of neurons are used sequentially, we speak of DNNs [20].

Most DNN architectures consist of a convolutional neural network (CNN) layer, followed by either a feedforward neural network (FNN) layer or a recurrent neural network (RNN) layer. Unlike the output from an RNN neuron, which is fed back into the same layer, the output from an FNN neuron is only connected to the next layer. CNNs handle variable input dimensions quite well and are mainly used for feature extraction for the RNN or FNN layer, which, combined with a prior CNN, output a better generalization than if fed with raw sensor data [22]. However, FNNs only work well with data of the same input dimensions, and RNNs only work with a fixed number of streams. As a result, the widely used CNN-RNN-FNN combinations do not work with varying input dimensions. This means that if data collection from one sensor stops, the movement type cannot be classified by the DNN that was trained on multiple input dimensions. In order to save battery life in smartphones during long-term recordings, it is often desirable to temporarily disable certain sensors or to vary the sampling rate of sensors, which results in changing the input dimensions for the DNN.

When a participant is sitting for an extended period, disabling the gyroscope sensor can conserve battery life. This is because the rotational position is unlikely to change significantly without significant acceleration changes unless the person is in an aircraft and the gravitational acceleration is being compensated for in the data. In order to determine when the activity type changes, it is sufficient to use a low recording frequency. This means that it is possible to deactivate the gyroscope and magnetometer and lower the accelerometer recording frequency. To determine when the activity type changes, a very low recording frequency suffices, so it is desirable to deactivate the gyroscope and magnetometer and lower the accelerometer recording frequency significantly. Dummy data can be generated to compensate for missing data in order to maintain the accuracy of the trained CNN-FNN-RNN model [23]. However, this approach can result in a loss of accuracy in classification. Another solution is to insert a global pooling layer, but this also leads to a reduction in accuracy. This, however, leads to accuracy loss in classification. Another solution is to insert a global pooling layer [24], but this also leads to a reduction in accuracy.

Previous publications on accelerometry-based movement recognition have shown great success but significant limitations. Ordóñez and Roggen [15] presented a deep-CNN-based framework, which they tested against models such as decision tree, random forest, and support vector machines. Trained and then tested on a data set, the accuracy reached up to 86.7%. The authors then analyzed which component of the data had the biggest impact on classification accuracy and determined this to be changes in acceleration, which is in line with our own results.

Wang et al [11] offer a comprehensive survey of recent advancements in activity recognition and associated methodologies. Their work sheds light on the various strengths and weaknesses of deep learning models when it comes to

activity classification. Although most models perform accurately on their trained data [25], significant limitations remain. First, the lack of extensive, labeled accelerometry data sets limits their efficacy. Second, the generalization capabilities of models need improvement. Third, models struggle with sensor noise and input variability, highlighting a need for greater robustness. Our algorithms aim to address these issues, working to mitigate the associated limitations and enhance overall model performance. To achieve this, we build upon previous research by incorporating and improving upon their methodologies while also introducing our own additional data set for algorithm training.

Malekzadeh et al [26] proposed a new model, which tries to counteract the aforementioned shortcomings by introducing a *dimension-adaptive pooling* (DAP) layer, which makes DNNs robust to changes in not only sampling rates but also dimensional changes of the data due to varying sensor availability.

The authors also introduced a *dimension-adaptive training* layer, and combined it with the classical CNN-FNN-RNN approach and the DAP layer. They claim that dimension-adaptive neural architecture (DANA) can prevent losses in classification accuracy, even under varying sensor availability and temporal sampling rate changes. This model was tested on 4 publicly available data sets, including the MotionSense [27] data set, which consists of accelerometer data from 24 students at Queen Mary University of London.

Our goal was to not only implement this model into our own DNN, but also to improve upon it and validate it using our own data. The robustness of the DANA model is very promising, making it a valuable addition to our research.

Methods

Ethical Considerations

According to the guidelines stated on the Ethics Commission page of the University of Bern's Faculty of Human Sciences, no ethics committee approval was required for this research.

This conclusion is based on the fact that all data was collected with participants' informed consent, the data collection was conducted anonymously, and the research activities only involved non-hazardous tasks such as standing, sitting, walking, and ascending or descending stairs. No personal data was collected.

Training Data

The data used for the initial training of the neural network was gathered from the MotionSense Github repository. These data consist of accelerometer and gyroscope readings from an iPhone 6s (Apple Inc), collected at a frequency of 50 Hz by 24 participants who followed a set of actions on the campus of Queen Mary University of London. These actions included ascending or descending stairs, sitting, walking, standing, and jogging (Figure 1). The data recorded gravity, acceleration, rotation, and attitude on 3 axes.

After conducting a principal component analysis, we found that the X, Y, and Z acceleration and rotational changes were the most predictive factors in classifying the participant's behavior (Figure 2). Therefore, only these 6 values were used in the training of the algorithm. As a result, our app only records these 6 values, which are then used for further analysis.

To gather more data and validate our model, we set up our own course of action on the campus of the Centre for Sports Science at the University of Bern, modeled after the course used at Queen Mary University. A total of 68 participants (aged 21-59, median 26, SD 3.2 years), who were students and employees of the University of Bern, completed the course while our *HumanActivityRecorder* Android app (Multimedia Appendix 1) was running and collecting data. All participants were fully informed about the task and gave their consent for the data collection.

The course consisted of approximately 300 seconds of walking, jogging, sitting, and walking up and down stairs and standing still (Figure 3). All participants completed all segments of the course, and the corresponding data segments were manually labeled for use in training the models.

Figure 1. Course for accelerometer data collection on the campus of the Queen Mary University of London for the MotionSense data set; graph from Malekzadeh et al [26].

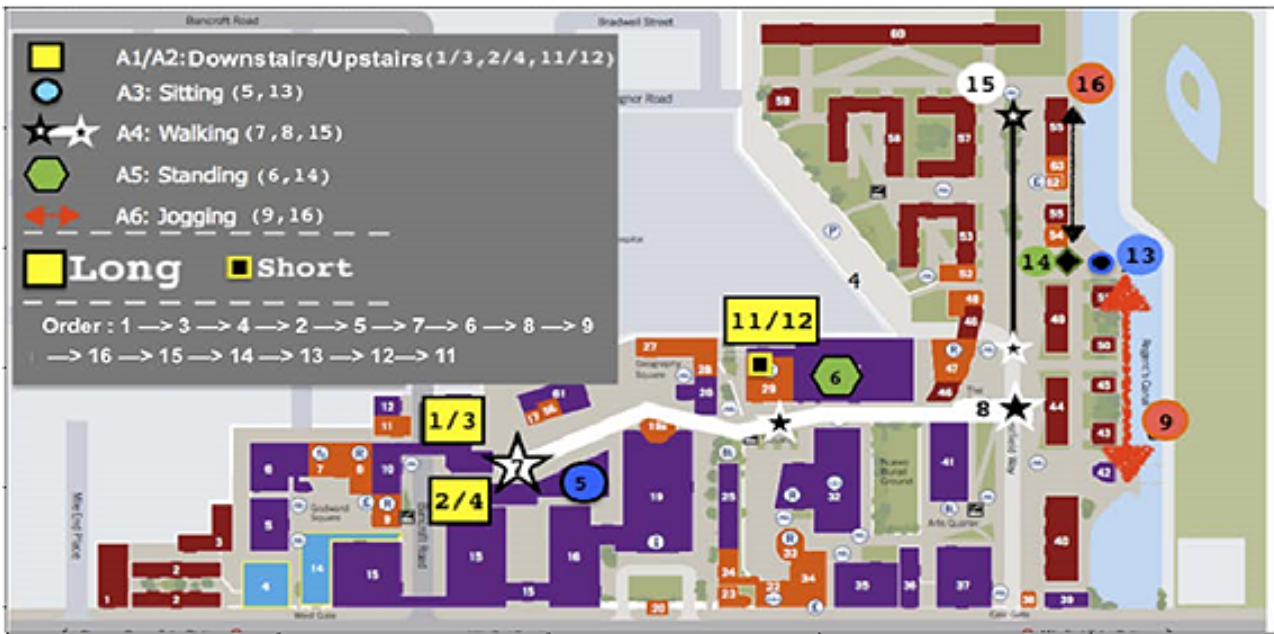


Figure 2. Data example of the MotionSense data set. Note that some values do not change significantly when normalized over the course of recording and are therefore of lesser interest for the prediction of behavior.

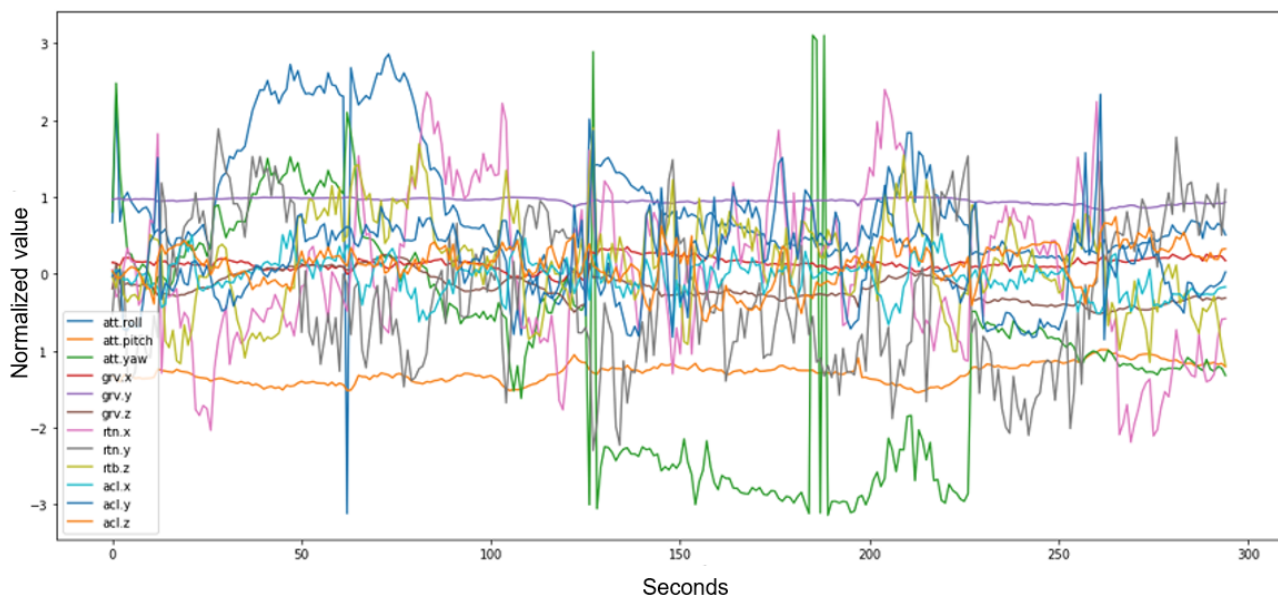


Figure 3. Course on the premises of the University of Bern. Participants followed the indicated path, starting walking, followed by jogging, sitting, ascending stairs, standing, and descending stairs. Completion took an average of approximately 300 seconds.

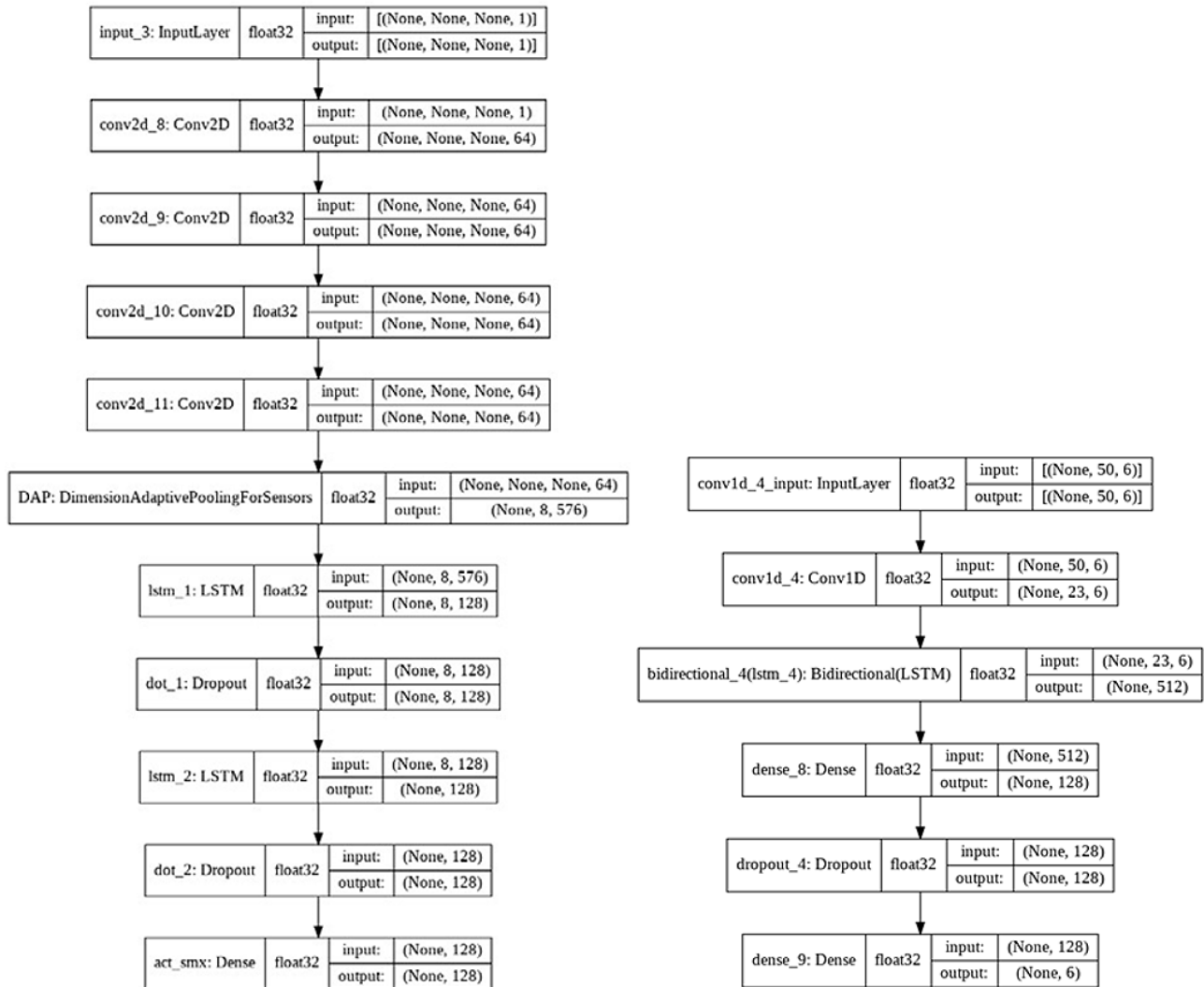


The participants completed the course in 2 groups with different instructions. Group 1 ($n=29$, median age 26, SD 5.2 years) was instructed to wear the smartphone in their preferred manner. Group 2 ($n=39$, median age 27, SD 4.7 years) wore the smartphone in the right front trousers' pocket, with the display facing toward the body and the top of the phone pointing down while standing. This placement is consistent with the data collection method used for the *MotionSense* data set, as discussed above. It was found that the orientation of the smartphone has a significant impact on the performance of the model. To ensure consistency and comparability between the data sets, our algorithm was trained on the data of group 2, as wearing the smartphone in an individually preferred manner (group 1) resulted in significantly worse performance in classification accuracy. For a detailed comparison of classification accuracy between groups 1 and 2, please refer to [Multimedia Appendix 2](#).

App

The accelerometer and gyroscope data were collected using our custom-made *HumanActivityRecorder* Android app, which was developed using Android Studio 4.1 with Java 1.8.0_271 ([Figure 4](#)). The app records accelerometer and gyroscope data at a sampling rate of 50 Hz and is publicly available on the Google Play Store as version 13 of the *HumanActivityRecorder* app. The accelerometer data are recorded in the x-, y-, and z-axes, while the gyroscope data consist of rotation around these axes (roll, pitch, and yaw) at the same frequency. The data are then automatically sent to a server and can be downloaded as a CSV or JSON file. The source code is available on Github [28]. The app is compatible with Android 5.0 and later versions. We used an Honor View 20 smartphone for data collection to ensure consistency in recording. Only 1 device was used.

Figure 4. Comparison of the models used in our study. The dimension-adaptive neural architecture (DANA) model, consists of several additional layers, which we found did not improve the classification of our data. Note that in our simplified model, the dimension-adaptive pooling (DAP) layer has been omitted as well, since our data are dimensionally consistent. LSTM: Long short-term memory.



Recording

Before beginning the data collection process, the participants were asked for their name, age, and consent. The data collection paradigm was explained to them and demonstrated through a walk-through by the data collector. The participants then completed the course, which included walking, jogging, sitting, ascending and descending stairs, and standing still, while the app recorded their accelerometer and gyroscope data. After completing the course, the participants were given a chocolate bar as an incentive. The accelerometer data were processed and categorized using a Jupyter notebook script, which automates the workflow to ensure consistency in categorization. This script is part of our toolbox.

Deep Learning Model

We implemented a modified version of the DANA model proposed by Malekzadeh et al [19], which involved removing and modifying several layers. This modification was made after testing the model (trained and tested on *MotionSense* data) and finding that the omission of these layers did not noticeably decrease the model's performance.

It is important to note that in our simplified model, we removed the DAP layer as our input data are dimensionally consistent at the time of testing. To validate the models, we trained them both on the *MotionSense* data set and our own data set, as well as testing both combinations.

Results

Through a systematic variation of the number of nodes and layers, we determined that the best balance between accuracy and complexity is achieved with the described architecture. This architecture was determined based on the accuracy of the models in classifying movement types of the *MotionSense* data set when trained on the same data set. Interestingly, when we trained on the *MotionSense* data set and tested on our own data, our model performed better than DANA, yet still with room for improvement, at 63% vs 26%.

When trained on the same data set as the one they are tested on, both models performed well in classifying behavior. The DANA model achieved approximately 87% accuracy when trained and tested on the *MotionSense* data set and approximately 90% accuracy when trained and tested on our own data, depending

on the sampling rate (Figure 5). However, when trained on the MotionSense data set and tested on our own data, the accuracy of DANA drops to around 26%, also depending on the dimensionality of the input, while our model performs at around 63%, but much less robust against the dimensionality input (Figure 6). This still leaves room for improvement but shows

the comparatively high generalization ability of our model. It is important to note that neither the MotionSense data nor our own data include magnetometer data, which is why the DANA model performs poorly (at or near zero accuracy) when reduced to only magnetometer input. The graph includes this information for consistency.

Figure 5. Accuracy in classifying using the dimension-adaptive neural architecture (DANA) model (A) trained and tested on MotionSense data; (B) our model trained and tested on our data; (C) DANA trained on MotionSense and tested on our data; and (D) our model trained on our own data and tested on MotionSense data. Note that the dimensionality is varied here to showcase the robustness, and our model is impacted more strongly by a varied dimensionality input. Acc: accelerometer; Gyr: gyroscope; Mag: magnetometer.

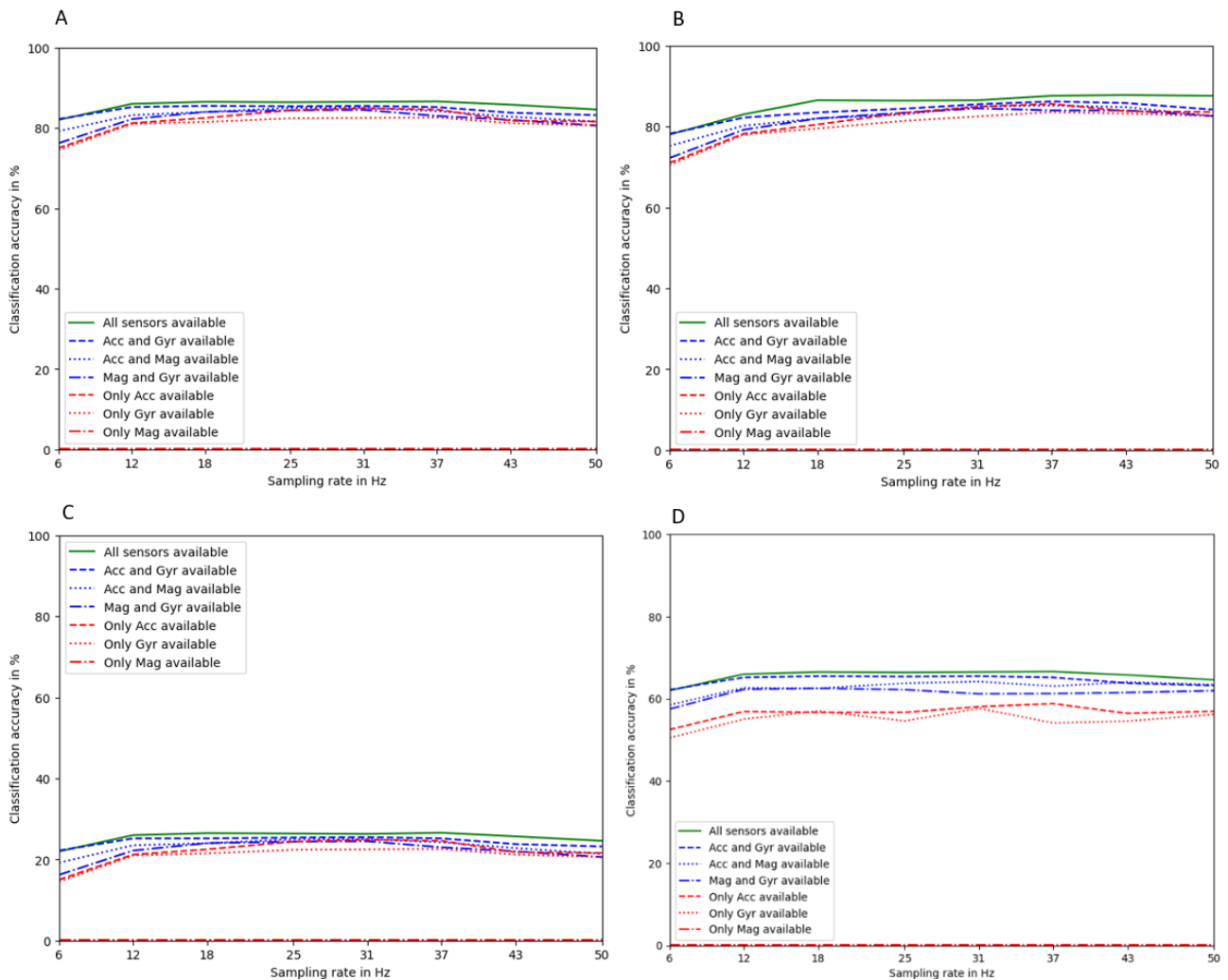
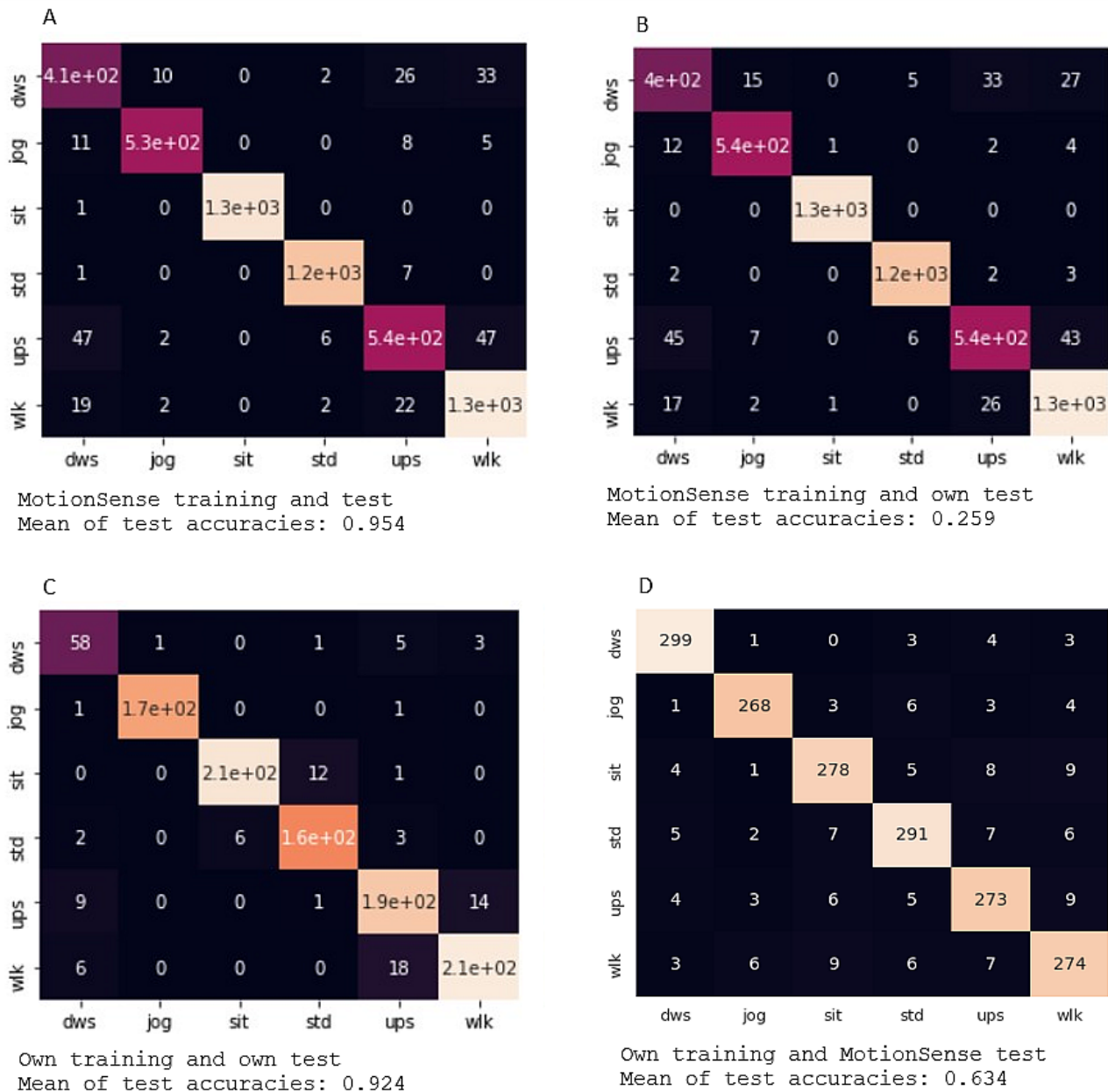


Figure 6. Confusion matrices of accuracy in classifying (A) using our own simplified model trained on MotionSense data tested on MotionSense data; (B) trained on MotionSense data and tested on own data; (C) trained and tested on our own data; and (D) trained on our own data and tested on MotionSense data. Note that dimensionality is not varied here as all sensors are available. dws: downstairs; jog: jogging; sit: sitting; std: standing; ups: upstairs; wlk: walking.



Our simplified model does not include the DAP layer and is less robust against input dimensional variance, as our input data dimensions did not vary. However, it is easily adaptable if desired. Despite this, our model outperforms the DANA model in terms of accuracy. When trained on the MotionSense data set and tested on it, our model achieved 95.4% accuracy. It was equally accurate when trained on our own data and tested on it, with 92.4% accuracy. However, when trained on the MotionSense data and tested on our own data, accuracy drops to 25.8%, but when trained on our data and tested on MotionSense, accuracy reached 63.4%.

Discussion

Conclusions

Both models included in our toolbox perform well when trained and tested on the same data set. However, they do not perform well when trained on one data set and tested on the other, as was the case in our study. This highlights the problem of the unavoidable part of overfitting the collected data to improve algorithm performance, although this is controlled for as far as possible. Despite this, both models (DANA and our own) performed similarly when trained on one data set and tested on the other. Our model is slightly more accurate, but the DANA model is more robust with regards to dimensional variance in the input. However, there is a significant difference in computing time when training the models. The DANA model, when trained

using Google Colab with CPU and GPU resources, took around 11 hours to train each time. On the other hand, our model can be trained in about 5 minutes with 100 epochs of training using only CPUs in Google Colab. Note that this estimation does not include hyperparameter testing.

Given the amount of data used to train the models, the results are surprisingly accurate. Commercial wearables, such as sports-oriented smartwatches, often have a function to display the user's current activity. However, these displayed activities are often incorrect, even for activities that seem obvious to the user. Considering these devices are widely available and sold to millions of people, we expected movement detection to be much more challenging, and our accuracy to be in the low 60% range.

While the accuracy of movement classification is very good, there is still room for improvement, which we plan to achieve by training the algorithm on additional data from diverse populations or environments. We recommend using the DANA model to classify behavior in data that have been gathered at different dimensions or with variable input dimensions. However, if the input type is consistent, we recommend our model as it is slightly more accurate and much easier to train. Both algorithms are available at our Github repository, along with the *HumanActivityRecorder* app and the scripts to process the data. In a future step, we plan to integrate both algorithms

into the app and evaluate their performance in a subsequent study.

Limitations

The orientation of the smartphone during recording has an impact on classification accuracy if the sample size is not large enough, as shown in our comparison of classification accuracy of groups 1 and 2 ([Multimedia Appendix 2](#)). However, if trained on large data sets with varying orientation, this effect disappears. For comparability, we based our model on the group with the same orientation as in the *MotionSense* data set. Accounting for orientation was outside the scope of our study. To address the impact of smartphone orientation on classification accuracy in medium-sized samples, an easy solution would be to incorporate an orientation recognition stage that detects the orientation of the smartphone and branches the data to models that have been individually trained on each orientation. This would ensure more accurate classification regardless of the smartphone orientation.

Authenticity

The results of the study are presented clearly, honestly, and without fabrication, falsification, or inappropriate data manipulation. The results of this study do not constitute endorsement by this Journal. This manuscript has not been published elsewhere, and it has not been submitted simultaneously for publication elsewhere.

Acknowledgments

This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

Data Availability

All data used are available [[28](#)].

Authors' Contributions

FW was the principal investigator, drafted the manuscript, and trained the algorithm; CN provided guidance for publishing.

Conflicts of Interest

None declared.

Multimedia Appendix 1

Screenshots of the Android app. From left to right: start screen, sociodemographics, and recording screen.

[[PNG File , 151 KB - ai_v2i1e42337_app1.png](#)]

Multimedia Appendix 2

Accuracy of the classification of our model (A) trained and tested on group 1 data; (B) trained on group 1 data and tested on MotionSense data; (C) trained and tested on group 2 data; and (D) trained on group 2 data and tested on MotionSense data. Group 1 was instructed to wear the smartphone wherever they preferred individually. Group 2 was instructed to wear it screen inside, top facing downward in the right trouser pocket, in line with data collection for the MotionSense data set, to ensure maximum comparability.

[[PNG File , 139 KB - ai_v2i1e42337_app2.png](#)]

References

1. Number of smartphone mobile network subscriptions worldwide from 2016 to 2022, with forecasts from 2023 to 2028. Statista. URL: <http://www.statista.com/statistics/330695/number-of-smartphone-users-worldwide> [accessed 2023-05-18]
2. Mobile Consumer Survey 2017: The UK cut. Deloitte. URL: <https://www.deloitte.co.uk/mobileuk2017/> [accessed 2023-05-18]

3. Tacconi C, Mellone S, Chiari L. Smartphone-Based Applications for Investigating Falls and Mobility. 2011 Presented at: Proceedings of the 5th International ICST Conference on Pervasive Computing Technologies for Healthcare; May 23-26, 2011; Dublin, Republic of Ireland. [doi: [10.4108/icst.pervasivehealth.2011.246060](https://doi.org/10.4108/icst.pervasivehealth.2011.246060)]
4. Mehta DD, Zañartu M, Feng SW, Cheyne HA, Hillman RE. Mobile Voice Health Monitoring Using a Wearable Accelerometer Sensor and a Smartphone Platform. *IEEE Trans. Biomed. Eng* 2012 Nov;59(11):3090-3096. [doi: [10.1109/tbme.2012.2207896](https://doi.org/10.1109/tbme.2012.2207896)]
5. Garcia-Ceja E, Osmani V, Mayora O. Automatic Stress Detection in Working Environments From Smartphones' Accelerometer Data: A First Step. *IEEE J. Biomed. Health Inform* 2016 Jul;20(4):1053-1060. [doi: [10.1109/jbhi.2015.2446195](https://doi.org/10.1109/jbhi.2015.2446195)]
6. Fino E, Mazzetti M. Monitoring healthy and disturbed sleep through smartphone applications: a review of experimental evidence. *Sleep Breath* 2019 Mar 23;23(1):13-24. [doi: [10.1007/s11325-018-1661-3](https://doi.org/10.1007/s11325-018-1661-3)] [Medline: [29687190](https://pubmed.ncbi.nlm.nih.gov/29687190/)]
7. Lau S, David K. Movement recognition using the accelerometer in smartphones. 2010 Presented at: 2010 Future Network & Mobile Summit; June 16-18, 2010; Florence, Italy.
8. Lee Y, Cho S. Activity Recognition Using Hierarchical Hidden Markov Models on a Smartphone with 3D Accelerometer. 2011 Presented at: HAIS 2011: Hybrid Artificial Intelligent Systems; September 22-24, 2011; Bilbao, Spain p. 460-467. [doi: [10.1007/978-3-642-21219-2_58](https://doi.org/10.1007/978-3-642-21219-2_58)]
9. Wannenburg J, Malekian R. Physical Activity Recognition From Smartphone Accelerometer Data for User Context Awareness Sensing. *IEEE Trans. Syst. Man Cybern, Syst* 2017 Dec;47(12):3142-3149. [doi: [10.1109/tsmc.2016.2562509](https://doi.org/10.1109/tsmc.2016.2562509)]
10. Case MA, Burwick HA, Volpp KG, Patel MS. Accuracy of smartphone applications and wearable devices for tracking physical activity data. *JAMA* 2015 Feb 10;313(6):625-626. [doi: [10.1001/jama.2014.17841](https://doi.org/10.1001/jama.2014.17841)] [Medline: [25668268](https://pubmed.ncbi.nlm.nih.gov/25668268/)]
11. Wang J, Chen Y, Hao S, Peng X, Hu L. Deep learning for sensor-based activity recognition: A survey. *Pattern Recognition Letters* 2019 Mar;119:3-11. [doi: [10.1016/j.patrec.2018.02.010](https://doi.org/10.1016/j.patrec.2018.02.010)]
12. Yang J, Nguyen M, San P, Li X, Krishnaswamy S. Deep Convolutional Neural Networks on Multichannel Time Series for Human Activity Recognition. 2015 Presented at: Proceedings of the 24th International Conference on Artificial Intelligence; July 25-31, 2015; Buenos Aires, Argentina.
13. Ronao CA, Cho S. Human activity recognition with smartphone sensors using deep learning neural networks. *Expert Systems with Applications* 2016 Oct;59:235-244. [doi: [10.1016/j.eswa.2016.04.032](https://doi.org/10.1016/j.eswa.2016.04.032)]
14. Ignatov A. Real-time human activity recognition from accelerometer data using Convolutional Neural Networks. *Applied Soft Computing* 2018 Jan;62:915-922. [doi: [10.1016/j.asoc.2017.09.027](https://doi.org/10.1016/j.asoc.2017.09.027)]
15. Ordóñez FJ, Roggen D. Deep Convolutional and LSTM Recurrent Neural Networks for Multimodal Wearable Activity Recognition. *Sensors (Basel)* 2016 Jan 18;16(1):115 [FREE Full text] [doi: [10.3390/s16010115](https://doi.org/10.3390/s16010115)] [Medline: [26797612](https://pubmed.ncbi.nlm.nih.gov/26797612/)]
16. Zhao Y, Yang R, Chevalier G, Xu X, Zhang Z. Deep Residual Bidir-LSTM for Human Activity Recognition Using Wearable Sensors. *Mathematical Problems in Engineering* 2018 Dec 30;2018:1-13. [doi: [10.1155/2018/7316954](https://doi.org/10.1155/2018/7316954)]
17. Yao S, Hu S, Zhao Y, Zhang A, Abdelzaher T. DeepSense: A Unified Deep Learning Framework for Time-Series Mobile Sensing Data Processing. 2017 Presented at: Proceedings of the 26th International Conference on World Wide Web; April 3--7, 2017; Perth, Australia. [doi: [10.1145/3038912.3052577](https://doi.org/10.1145/3038912.3052577)]
18. Jeyakumar J, Lai L, Suda N, Srivastava M. SenseHAR: a robust virtual activity sensor for smartphones and wearables. 2019 Presented at: Proceedings of the 17th Conference on Embedded Networked Sensor Systems; November 10-13, 2019; New York, USA p. 15-28. [doi: [10.1145/3356250.3360032](https://doi.org/10.1145/3356250.3360032)]
19. Malekzadeh M, Clegg RG, Cavallaro A, Haddadi H. Privacy and utility preserving sensor-data transformations. *Pervasive and Mobile Computing* 2020 Mar;63:101132. [doi: [10.1016/j.pmcj.2020.101132](https://doi.org/10.1016/j.pmcj.2020.101132)]
20. Schmidhuber J. Deep learning in neural networks: an overview. *Neural Netw* 2015 Jan;61:85-117. [doi: [10.1016/j.neunet.2014.09.003](https://doi.org/10.1016/j.neunet.2014.09.003)] [Medline: [25462637](https://pubmed.ncbi.nlm.nih.gov/25462637/)]
21. Goodfellow I, Bengio Y, Courville A. *Deep Learning*. Cambridge, Massachusetts, USA: MIT press; 2016.
22. Bengio Y, Courville A, Vincent P. Representation learning: a review and new perspectives. *IEEE Trans Pattern Anal Mach Intell* 2013 Aug;35(8):1798-1828. [doi: [10.1109/TPAMI.2013.50](https://doi.org/10.1109/TPAMI.2013.50)] [Medline: [23787338](https://pubmed.ncbi.nlm.nih.gov/23787338/)]
23. Lee JA, Gill J. Missing value imputation for physical activity data measured by accelerometer. *Stat Methods Med Res* 2018 Feb 17;27(2):490-506. [doi: [10.1177/0962280216633248](https://doi.org/10.1177/0962280216633248)] [Medline: [26994215](https://pubmed.ncbi.nlm.nih.gov/26994215/)]
24. Lin M, Chen Q, Yan S. Network In Network. *arXiv* 2014:1-10 [FREE Full text] [doi: [10.48550/arXiv.1312.4400](https://doi.org/10.48550/arXiv.1312.4400)]
25. Islam MM, Nooruddin S, Karray F, Muhammad G. Human activity recognition using tools of convolutional neural networks: A state of the art review, data sets, challenges, and future prospects. *Comput Biol Med* 2022 Oct;149:106060. [doi: [10.1016/j.compbio.2022.106060](https://doi.org/10.1016/j.compbio.2022.106060)] [Medline: [36084382](https://pubmed.ncbi.nlm.nih.gov/36084382/)]
26. Malekzadeh M, Clegg R, Cavallaro A, Haddadi H. DANA. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol* 2021 Sep 14;5(3):1-27. [doi: [10.1145/3478074](https://doi.org/10.1145/3478074)]
27. MotionSense dataset. GitHub. URL: <https://github.com/mmalekzadeh/motion-sense> [accessed 2023-05-18]
28. HumanActivityRecorder. GitHub. URL: <https://github.com/FluWieland/HumanActivityRecorder> [accessed 2023-05-19]

Abbreviations

CNN: convolutional neural network
DANA: dimension-adaptive neural architecture
DAP: dimension-adaptive pooling
DNN: deep neural network
FNN: feedforward neural network
RNN: recurrent neural network

Edited by K El Emam, B Malin; submitted 21.09.22; peer-reviewed by H Li, G Lim, SAH Aqajari, Y Wang; comments to author 21.12.22; revised version received 28.02.23; accepted 22.04.23; published 08.06.23.

Please cite as:

Wieland F, Nigg C

A Trainable Open-Source Machine Learning Accelerometer Activity Recognition Toolbox: Deep Learning Approach

JMIR AI 2023;2:e42337

URL: <https://ai.jmir.org/2023/1/e42337>

doi: [10.2196/42337](https://doi.org/10.2196/42337)

PMID: [38875548](https://pubmed.ncbi.nlm.nih.gov/38875548/)

©Fluri Wieland, Claudio Nigg. Originally published in JMIR AI (<https://ai.jmir.org>), 08.06.2023. This is an open-access article distributed under the terms of the Creative Commons Attribution License (<https://creativecommons.org/licenses/by/4.0/>), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR AI, is properly cited. The complete bibliographic information, a link to the original publication on <https://www.ai.jmir.org/>, as well as this copyright and license information must be included.

Original Paper

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

Tharshini Jeyakumar¹, BSc, MHI; Sarah Younus¹, MPH; Melody Zhang¹, MA; Megan Clare², HBSc; Rebecca Charow^{1,3}, MSc; Inaara Karsan^{1,3}, BSc; Azra Dhalla⁴, MBA; Dalia Al-Mouaswas², HBSc; Jillian Scandiffio¹, MSc; Justin Aling⁵; Mohammad Salhia², MEd; Nadim Lalani⁴, BA; Scott Overholt⁵; David Wiljer^{1,3,6,7}, PhD

¹University Health Network, Toronto, ON, Canada

²Michener Institute of Education, University Health Network, Toronto, ON, Canada

³Institute of Health Policy, Management and Evaluation, Dalla Lana School of Public Health, University of Toronto, Toronto, ON, Canada

⁴Vector Institute, Toronto, ON, Canada

⁵Patient Partner Program, University Health Network, Toronto, ON, Canada

⁶Faculty of Medicine, University of Toronto, Toronto, ON, Canada

⁷Office of Education, Centre for Addiction and Mental Health, Toronto, ON, Canada

Corresponding Author:

David Wiljer, PhD

University Health Network

190 Elizabeth Street R. Fraser Elliott Building RFE 3S-441

Toronto, ON, M5G 2C4

Canada

Phone: 1 416 340 4800 ext 6322

Email: David.wiljer@uhn.ca

Abstract

Background: As new technologies emerge, there is a significant shift in the way care is delivered on a global scale. Artificial intelligence (AI) technologies have been rapidly and inexorably used to optimize patient outcomes, reduce health system costs, improve workflow efficiency, and enhance population health. Despite the widespread adoption of AI technologies, the literature on patient engagement and their perspectives on how AI will affect clinical care is scarce. Minimal patient engagement can limit the optimization of these novel technologies and contribute to suboptimal use in care settings.

Objective: We aimed to explore patients' views on what skills they believe health care professionals should have in preparation for this AI-enabled future and how we can better engage patients when adopting and deploying AI technologies in health care settings.

Methods: Semistructured interviews were conducted from August 2020 to December 2021 with 12 individuals who were a patient in any Canadian health care setting. Interviews were conducted until thematic saturation occurred. A thematic analysis approach outlined by Braun and Clarke was used to inductively analyze the data and identify overarching themes.

Results: Among the 12 patients interviewed, 8 (67%) were from urban settings and 4 (33%) were from rural settings. A majority of the participants were very comfortable with technology (n=6, 50%) and somewhat familiar with AI (n=7, 58%). In total, 3 themes emerged: cultivating patients' trust, fostering patient engagement, and establishing data governance and validation of AI technologies.

Conclusions: With the rapid surge of AI solutions, there is a critical need to understand patient values in advancing the quality of care and contributing to an equitable health system. Our study demonstrated that health care professionals play a synergetic role in the future of AI and digital technologies. Patient engagement is vital in addressing underlying health inequities and fostering an optimal care experience. Future research is warranted to understand and capture the diverse perspectives of patients with various racial, ethnic, and socioeconomic backgrounds.

KEYWORDS

artificial intelligence; patient; education; attitude; health data; adoption; health equity; patient engagement

Introduction

Background

Artificial intelligence (AI) technologies are being rapidly adopted and implemented in health care settings to augment clinical decisions and the delivery of patient-centered care [1]. The use of AI applications presents a paradigm shift in health care and serves as a positive enabler for achieving the quintuple aims of health care [2]. In particular, AI applications have the potential to further integrate health equity and patient activation to ameliorate siloed and biased care, as advocated by the National Academy of Medicine [2,3]. Fostering a patient-centered culture that considers health equity entails continued partnerships with patients and encourages them to be coleaders of change within the clinical ecosystem [2]. This shift must emerge from the leadership and organizational levels and should include both a commitment to and development of strategic priorities, which include patient and family engaged care [2]. For instance, the Canadian Institute for Advanced Research urges the need for a collaborative and integrative effort to establish an AI for Health strategy to accelerate the adoption and scaling of AI-enabled technologies to provide compassionate and safe care [4]. The Canadian Institute for Advanced Research highlights the importance of including patient perspectives in the development, implementation, and evaluation of AI initiatives [4]. A few studies have reported that a co-design approach engaging patients and the public during the development process could enhance the accuracy, equity, and transparency of AI models [5-7]. Patients are key beneficiaries in the adoption and implementation of AI technologies in clinical settings; thus, engaging patients allows for diversity in perspectives, and their values and needs are included [8,9].

Importance of Fostering Patient Engagement

Patient engagement is defined as an individual's active involvement in the care decision-making process and collaboration with key stakeholders to build an equitable and sustainable health system [10,11]. Understanding patient perceptions is an initial step in fostering patient engagement and ensuring the responsible and safe use of these novel technologies in clinical care settings [8]. A recent survey conducted by the Biron Health Group in Quebec indicated that many residents were in favor of using AI technologies to address health system issues and optimize clinical innovations [12]. The study showed that 63% agreed that AI could prevent adverse outcomes, while 40% believed that it could be used to augment clinicians' expertise and lead to profound changes in care [12]. Many papers focused on patient perspectives of AI in various medical specialties, such as cardiology, dermatology, and radiology, and how they conceptualize AI technology in health care [13-21]. Although there are several studies focused on understanding patient perspectives in relation to specific AI technologies, patients need to be engaged at different stages of the AI implementation process [9,22-24]. The long-term

sustainability of AI technologies in clinical environments vastly relies on patient acceptance, which is influenced by their knowledge and perception of opportunities as well as risks associated with using AI solutions [15].

Despite the positive views on the potential of clinical applications of AI and the promise of AI, there are many fears and misconceptions that remain. A few studies have shown that patients expressed concern regarding the use of personal health records for profit or being distorted by hackers, as this could have an impact on their employment or insurance coverage [15,25]. Balthazar et al [25] contended that even when patients have an in-depth understanding and thoughts on the appropriate use of their personal health information, they may not be able to understand the foundational concepts of machine learning models to make predictions or discern the difference between terms such as privacy and confidentiality. Another significant concern noted in the literature is the systemic bias that can potentially be embedded in AI models and that can stigmatize or marginalize certain populations [7,8,25]. Patients' perspectives on AI may differ based on their socioeconomic status, ethnicity, and vulnerability [25]. Furthermore, patient engagement helps to cocreate the health care system, address the underlying social determinants of health [2,26], and ultimately democratize access to AI innovations [5]. Thus, minimizing the consequences and concerns of AI technologies is pivotal in facilitating trust and ensuring the successful adoption of these tools in clinical practice.

Establishing patient trust becomes increasingly difficult in a rapidly evolving digital space with complex and less-transparent AI technologies [8]. Studies have asserted that even though AI can empower patients, the lack of transparency and explanation of processes owing to the *black box* phenomenon could diminish patients' trust if the model is not reflective of current evidence, is biased, or is erroneous [27-30]. Notwithstanding the high accuracy and advancements in AI technologies, patients value human judgment when making care decisions [31]. Empathy, compassion, and trust play a significant role in forming the basis for augmenting patient-centered care and ensuring the sustainability of AI innovations [27,32]. It is vital for care providers to actively engage patients when making care decisions and foster a therapeutic relationship [32]. Kerasidou [32] highlighted that patients preferred to interact with health care professionals (HCPs) who both have clinical expertise and provide empathetic and compassionate care. An interpersonal care model allows HCPs to better understand and address individual needs and to build patients' trust [7,32]. In addition, the literature emphasizes the importance of public perception and literacy in fostering trust and removing any potential misconceptions regarding AI [28]. Esmaeilzadeh et al [27] advocated for patient education to ensure that patients are prepared to make informed decisions and communicate effectively with their care providers. The authors underlined the importance of patients being active partners during the

adoption and integration of AI innovations in their care [27]. Thus, patient engagement helps diminish the gap between patients' expectations of AI technologies and their experiences with care providers [33].

Current Landscape

Cutting-edge technologies such as AI are poised to transform the health care system, as we slowly shift to a new revolution in the next era [20]. This shift is facilitated through medical education; however, there are gaps in its implementation across all levels of medical education. This includes the lack of standardization, varying levels of AI literacy among faculty, and limited infrastructure for embedding AI concepts within existing curricula [34]. There is a need for medical education to go beyond medical informatics and machine learning, enabling HCPs to operationalize these novel tools at the point of care [34]. Despite the use of AI to accelerate innovations in patient care and the need for patient voices, there is limited literature on patient engagement and their perceptions of how AI will affect care delivery, thus ensuring AI technologies are aptly integrated within the clinical environment and cultivating patient trust [9]. To put the needs of patients first in creating a healthier world using AI, the objective of this study was to elucidate patients' perceptions of what skills they believe HCPs should have in preparation for this AI-enabled future and how we can better engage patients when adopting and deploying AI technologies in health care.

Methods

Study Design

A qualitative study design was used to elicit participants' perceptions of the adoption and implementation of AI within the health care ecosystem.

Ethics Approval

This study was approved by the University Health Network Research Ethics Board (ID:20-6148.2).

Study Participants

A maximum purposive sampling approach was used to ensure that the participants represented various comfort levels with AI technology and contexts in which they received care. It was also used to gain insights into the diverse perspectives that should be considered when adopting and deploying AI technologies in clinical settings. Purposive sampling enables researchers to identify and select participants based on their ability to yield relevant information about a particular phenomenon [35,36]. Participants were recruited from a national group of approximately 25 patients via email invitations sent on behalf of the research team by education committee members of Canada Health Infoway. Participants who consented to participate in the interviews were asked to inform individuals within their networks via a snowball sampling approach [37]. The snowball sampling method was used to recruit additional participants, who may add valuable perspectives to the study and enable an in-depth understanding of the phenomenon. Individuals were eligible to participate if they were patients at

any Canadian medical center (acute or long-term) and were able to provide informed consent.

Data Collection

Semistructured interviews were conducted with patients on the web via Microsoft Teams, in following COVID-19 pandemic social distancing measures. An instructional designer and research associates who have experience in qualitative research methods conducted the interviews. In addition, the interviewers have formal education in health informatics (TJ), public health (SY), educational technology (MC), and educational and counseling psychology (MZ). A semistructured interview guide consisting of 13 open-ended questions was used to guide discussions (Multimedia Appendix 1). The interviewers probed participants when necessary to further explore and understand salient ideas. The participants' level of comfort in sharing their perceptions and experiences determined the length of the interview. The interviews lasted approximately 17 to 48 minutes. The interviews were conducted until the researchers felt that no new ideas emerged and data saturation was achieved. Participants were offered an honorarium of CAD \$50 (US \$37.32) in the form of e-gift cards. Verbal informed consent was obtained before conducting interviews. All interviews were digitally audio-recorded, professionally transcribed, and deidentified. The transcripts were reviewed for accuracy by a research associate.

Data Analysis

Reflexivity is crucial in qualitative research, as it enables researchers to position themselves and reflect on the biases, values, and experiences that they bring [38,39]. Recognizing the researchers' perspectives and positionality, research rigor was asserted by providing a reflexive stance in the research process, including different viewpoints from the team. Seven members of the core research team participated in the coding and analytic process, including 4 research associates from the digital education department at a large multisite academic health sciences center (TJ, SY, MZ, and SB), instructional designer (MC), 2 patient partners (JA and SO), and a senior investigator (DW, a PhD education researcher). This enabled a rigorous interpretation and analysis of the findings. A systematic process outlined by Braun and Clarke [40] was used to inductively analyze the data. Two research associates (TJ and SY) independently analyzed the first 3 transcripts from an exploratory lens and developed an initial coding structure. Each of the remaining transcripts were coded independently by two study team members (TJ and MC, MZ or SB). New data were constantly compared with the existing data, thus resulting in iterative refinement of the coding structure and the structuring of further data collection. Iterative discussions with the research team helped contextualize the overarching themes and resolve disagreements. The senior investigator (DW) on the team reviewed all themes and provided additional input when consensus could not be reached. Two patient partners who were part of the study team (JA and SO) reviewed the themes, which allowed for triangulation of the data from various perspectives. Data were analyzed for emerging themes using NVivo version 12 (QSR International), a qualitative data analysis software program. The rigor and quality of thematic analysis were

evaluated using a 20-question evaluation tool [41]. The team also maintained a record of each team member's coding, notes from meetings, and different versions of the coding structure. This review enhanced the credibility and trustworthiness of the findings. Furthermore, an intercoder agreement was established using NVivo 12 to ensure transparency and rigor of the data.

Results

Overview

In total, 12 interviews were conducted between August 2021 and December 2021. Of the 12 participants, 10 (83%) were

females, and 2 (17%) were males. Table 1 shows the characteristics of the study participants. The average length of the interviews was 30 minutes. Most participants were very comfortable with the technology and somewhat familiar with AI. Thematic analysis of the data yielded three major themes, each with several subthemes (Table 2): (1) cultivating patients' trust, (2) fostering patient engagement, and (3) establishing data governance and validation of AI technologies.

Table 1. Participant characteristics (N=12).

Characteristic	Value, n (%)
Demographics	
Age (years)	
Young adult (18-40)	0 (0)
Middle age (40-60)	7 (58)
Senior (≥60)	5 (42)
Sex	
Male	2 (17)
Female	10 (83)
Location	
Urban	8 (67)
Rural	4 (33)
Comfort with technology	
Not at all comfortable	3 (25)
Somewhat comfortable	3 (25)
Very comfortable	6 (50)
Familiarity with AI^a	
Not at all familiar	3 (25)
Somewhat familiar	7 (58)
Very familiar	2 (17)
AI information source	
Family and friends	4 (33)
Career	5 (42)
Scholarly articles	2 (17)
Non-peer-reviewed articles	3 (25)
Social media	2 (17)
Other	3 (25)
Medical care	
Frequency of visiting an HCP^b	
Once a year	2 (17)
Fewer than 4 times a year	2 (17)
4 to 6 times a year	8 (66)
Type of HCP	
Cardiologist	3 (25)
General practitioner	10 (83)
Ophthalmologist	3 (25)
Physiotherapist	3 (25)
Surgeon	3 (25)
Other	8 (67)

^aAI: artificial intelligence.

^bHCP: health care professional.

Table 2. Summary of key themes.

Theme and quote	Significance
Theme 1: cultivating patients' trust	
Subtheme: providing safe and compassionate care	
<ul style="list-style-type: none"> “I would feel comfortable as long as I still had a voice. And they listen to the voice, OK, as opposed to the data... I mean, if I trust my health care provider and they're thorough and reliable, I would go along with it.” [ID 8] “I mean, I think I would worry about us totally removing the human part of this. That compassion and connection with a person who understands your health condition is really important... I would like a person who understands the question that I'm asking. So, I think it's making sure that we don't undervalue the importance of connection to other human beings, especially when we're talking about health care and the fears and anxieties that come up, about our health, so that we have someone who can not only answer our questions, but understand our fears and worries...” [ID 9] 	<ul style="list-style-type: none"> Transparent communication and acknowledging patient concerns and needs are imperative in fostering patients' trust. Most importantly, participants seemed critical of the use of digital technologies and their impact on therapeutic relationships. Compassion was identified as crucial in achieving patient-centered care; and ensuring the presence of technology does not encumber the human and relational aspects of a patient-provider relationship.
Subtheme: achieving transparency in care decisions	
<ul style="list-style-type: none"> “I want to know for sure like that it's a legitimate app that it's recommended by like major hospitals and those sorts of things, because right now everybody's making apps and it's very hard to tell what's real and what's not, especially at my age. I find my generation, my husband, we're much less trusting and we get confused, like the example of that bot that I was very unhappy with the bot being there [instead of a person]. But I would also be good if, let's say there are apps that it was overseeing. So, with a hospital, those sorts of things, like I really would like proof. And if it was dealing with my physician, well, then having her backing that would make me feel more comfortable using the app as well.” [ID 2] 	<ul style="list-style-type: none"> Given the rapid proliferation of digital technologies for patient care, participants stressed the need to be governed or regulated by the organization for the privacy and legitimacy of the app.
Theme 2: fostering patient engagement	
Subtheme: enabling patients to be coleaders in their care	
<ul style="list-style-type: none"> “The only thing I would say at the outset would be it's the machine that is running the process and I would want to be assured that the patient's feelings and voice would still be heard. Because there are things that, you know, there are things maybe ninety-nine percent going one way, but there is still that one percent that maybe the patient feels. Maybe there's other things going on with that patient that would come out in a meeting with a doctor.” [ID 8] “I think it's important that we as patients are as involved in our care as possible. I would like to expect that my GP would engage me in the decision-making about my care, even if an algorithm directed him to do something or not do something. I think that's an important aspect of communication.” [ID 4] 	<ul style="list-style-type: none"> Participants highlighted the importance of HCPs^a engaging them during the clinical decision-making process and providing an opportunity for them to share their thoughts and perspectives.
Subtheme: increasing confidence among patients	

Theme and quote	Significance
<ul style="list-style-type: none"> “Some people will want to know a lot and some people will want to know less. But certainly, the overall importance of sharing on some level so that we can improve our systems, I think is critical, but how do we do it safely? And if we can explain that to people in a way that gives them confidence and that they know their information will not be released to the wrong people in an identifying way, that’s important, but it doesn’t obviate the risk completely. So, I still think that you know, people need to at least have the opportunity to understand that this is a really complicated and important decision to make... how could that information be used in ways that are contrary to your best in financial health or otherwise?” [ID 4] “...a health professional who can also help me and guide me if there’s something that I don’t understand, or I’m missing a piece of this puzzle. So, a coach and educator. Yeah, someone who’s got my back with the AI as well. So again, I just think we can’t lose sight of that human touch and how we learn and digest and understand information. It’s not just a transaction.” [ID 9] 	<ul style="list-style-type: none"> When using technologies at the point of care, HCPs need to explain to patients the benefits and risks associated with it; thus, enabling them to be informed and understand how decisions are made. One participant emphasized that the patient-provider interaction is not a transaction as the technology can become a third player and the provider may neglect the compassionate aspect of the relationship. In addition, educating patients on the fundamentals of AI^b and other technologies can increase their confidence.

Theme 3: establishing data governance and validation of AI technologies

Subtheme: responsibility of data stewards

- “...I have strong objections to it being sold. I know the [organization] was making their data available to a private company at one point. And I know there are doctors in Ontario who feel that the health record is theirs, and they own it. And so, the information and it may be mine, but since they own the program that holds my data, they feel they have every right to sell it, and they do. So, I want more control over who gets to use it and why. And I mean, I think a lot of people would say, I’m fine for the public good, I’m fine with research that will benefit me, and people like me. But they draw the line at people making money from their personal data.” [ID 10]
- “I would like to know if there’s any third parties going to see it. My other concern... Say the insurance, I tested positive for breast cancer, and it was a genetic one, I’m going through that right now. What how having AI and data out there on a computer without being shared with insurance companies, which is more likely to happen than it is right now. So, yeah, I would want to know how my privacy is being respected. And any third parties involved and any changes I’d want to be updated and if there were changes and third parties were going to see it, I’d have the choice of letting them or completely removing all my information.” [ID 2]
- Participants expressed concerns regarding how their data will be used and who will have access to it. They identified the need to provide them with a choice to opt-in or opt-out of the secondary use of data.

Subtheme: quality assurance and validation of AI technologies

- “Then that becomes no different if there’s no oversight or no background or no warnings about them or disclaimers, then it becomes just the same as people Googling everything. So, I would want it to be a better tool and a somewhat regulated tool or something so that it’s actually endorsed by the medical community before it’s available, or at least obviously they’re not going to be able to control everything that’s available on the Internet. But at least there would be some education to the public that to use the tools that we endorse or use the tool endorsed by your hospital or your province or whatever, there would be some kind of oversight. That’s all I’m concerned about, that it just becomes the next version of Google.” [ID 7]
- Quality assurance and validation of AI technologies are pivotal in ensuring the confidentiality of patient data and protecting them against nefarious acts.

Subtheme: ensuring AI technologies used in clinical contexts are equitable and inclusive

Theme and quote	Significance
<ul style="list-style-type: none"> “Oh, yeah, definitely as a tool to assist physicians, I think it would be great. And I think that there are circumstances where the artificial intelligence tool might do a better job than the doctor. Because you know, a lot of people in health care are... people have preconceived notions about them, right. For instance, if somebody decides that you’re a hysterical woman, you won’t get the same care as you would if you had didn’t have that notation in your health record. And so, I think that with the use of artificial intelligence, it takes out some of the bias.” [ID 10] “I guess it really depends on who has actually set up the AI and what biases they have and what has actually been programmed into the system and if that’s actually missing data, just because of the bias and missing marginalized populations or people that don’t have a lot of money or are of a different race. And look, I just think there was something that I saw a while back about an app, you know, telling somebody had heart attack symptoms, and if it was male, it would say you should go to the hospital. But if it was female, it was like, oh, you don’t have a heart attack. You have I’m guessing this was a while ago, I’m guessing probably anxiety! So, there’s like [sex] differences, too. And so, I just wonder about the disparities that could be created, if it hasn’t been created with the people that it’s looking at.” [ID 5] 	<ul style="list-style-type: none"> One of the participants stated that AI could be an unbiased tool for HCPs to use in their care as it removes some of the preconceived perceptions that lead to further marginalization of certain groups. HCPs need to become adept in examining and acknowledging implicit biases to make informed decisions and prevent unintended consequences on patient care.

^aHCP: health care professional.

^bAI: artificial intelligence.

Theme 1: Cultivating Patients’ Trust

Providing Safe and Compassionate Care

Most participants believed that trust is fundamental to ensuring that AI technologies are successfully integrated into clinical care settings. They would be comfortable using an AI-based application if they knew it was coming from a trusted source such as their health care provider. However, they also mentioned that they would feel uncomfortable if they did not have the opportunity to discuss the technology with their health care provider or did not have a follow-up conversation with them:

I would feel comfortable as long as I still had a voice. And they listen to the voice, OK, as opposed to the data...I mean, if I trust my health care provider and they’re thorough and reliable, I would go along with it. [ID 8]

Using this technology in conjunction with the clinician’s expertise helps foster trust and ensures greater accountability. A few participants asserted that they would prefer their care provider to use their own knowledge and experience to make an informed decision and not solely based on the technology itself. As technologies are being integrated into clinical settings, patients do not want anything to change in the way they interact with their care provider or the way in which information is provided:

I mean, I think I would worry about us totally removing the human part of this. That compassion and connection with a person who understands your health condition is really important...I would like a person who understands the question that I’m asking. So, I think it’s making sure that we don’t undervalue

the importance of connection to other human beings, especially when we’re talking about health care and the fears and anxieties that come up, about our health, so that we have someone who can not only answer our questions, but understand our fears and worries. [ID 9]

Participants indicated that face-to-face interactions and the clinician’s presence are important for creating a safe space and maintaining trust. Participants commented that having a conversation with a clinician, as opposed to only interacting with the AI technology, provides support and reassurance, particularly when discussing sensitive health concerns such as mental health issues.

Achieving Transparency in Care Decisions

Participants would like clear communication from their HCPs on what applications and analytic health care tools are available and whether they are being used in their care. The participants expressed their desire for transparency in how physicians combined their judgment and technology to arrive at diagnoses and care decisions. Other participants noted that care providers did not have to understand the technical aspects of AI technology but needed to be confident in what they are prescribing and practicing to ensure that it is safe for patients.

Several participants also reported that care providers who willingly answered their questions or demonstrated ways to interact with the technology significantly increased their confidence levels in the technology. One participant mentioned that in comparison with providers who chose not to explain or demonstrate an AI technology, having an HCP explain what they did greatly boosted a patient’s positive perception of the technology and their comfort with it. Some participants also

preferred to see how physicians interacted with the technology and process they used to make clinical decisions. Furthermore, patients would prefer guidance on using health technologies and ascertaining what information is relevant to their own health care. One participant mentioned that they liked information on how the backend technology of an AI-enabled mobile application (app) was created. Regardless of the degree to which patients wanted to understand how an app works, they conveyed the need for any apps used to be vetted and recommended by their HCP:

I want to know for sure like that it's a legitimate app that it's recommended by like major hospitals and those sorts of things, because right now everybody's making apps and it's very hard to tell what's real and what's not, especially at my age. I find my generation, my husband, we're much less trusting and we get confused, like the example of that bot that I was very unhappy with the bot being there [instead of a person]. But I would also be good if, let's say there are apps that it was overseeing. So, with a hospital, those sorts of things, like I really would like proof. And if it was dealing with my physician, well, then having her backing that would make me feel more comfortable using the app as well. [ID 2]

Differences were found in the level of knowledge patients want to know about how AI technologies or apps work and the potential impacts on care decisions. However, all participants expressed the importance of transparency and communication in an app or provider's process for making care recommendations or decisions. Patients also want to be informed of the AI technologies that exist and whether they should be used in their care. Although there was a difference in the level of knowledge patients wanted their HCPs to have, the participants emphasized comfort in their recommendations and transparency.

Theme 2: Fostering Patient Engagement

Enabling Patients to Be Coleaders in Their Care

Enabling patients to become coleaders is vital when using digital technologies to inform care decisions. Participants asserted that it is important for health care organizations to actively listen to and understand the needs of the public:

The only thing I would say at the outset would be it's the machine that is running the process and I would want to be assured that the patient's feelings and voice would still be heard. Because there are things that, you know, there are things maybe ninety-nine percent going one way, but there is still that one percent that maybe the patient feels. Maybe there's other things going on with that patient that would come out in a meeting with a doctor. [ID 8]

Two participants specifically mentioned that they would like to be engaged and involved in the shared decision-making process, which also helps foster trust. For instance, if the AI application detects a concern, the patient would expect the care provider to have a discussion with them to identify the next steps:

I think it's important that we as patients are as involved in our care as possible. I would like to expect that my GP would engage me in the decision-making about my care, even if an algorithm directed him to do something or not do something, I think that's an important aspect of communication. [ID 4]

Participants reported that the integration of digital solutions as part of patient care is contingent upon the relationships they have established with their HCPs.

Increasing Confidence Among Patients

In the use of an AI app or technology, participants expressed the need for a log-in ID; a password; and an accessible, easy-to-use interface. They commented that having access to technology, such as being able to view the results on a cloud platform or digital patient profile, would be valuable and aid in their decision-making process. Furthermore, participants highlighted the need for patient education:

Some people will want to know a lot and some people will want to know less. But certainly, the overall importance of sharing on some level so that we can improve our systems, I think is critical, but how do we do it safely? And if we can explain that to people in a way that gives them confidence and that they know their information will not be released to the wrong people in an identifying way, that's important, but it doesn't obviate the risk completely. So, I still think that you know, people need to at least have the opportunity to understand that this is a really complicated and important decision to make...how could that information be used in ways that are contrary to your best in financial health or otherwise? [ID 4]

Although patients do not need to understand all details of their diagnosis, it is essential to provide them with relevant information at the right level. Participants reported that education helps increase awareness of existing AI technologies and how these technologies are used to augment patient care. Another participant stated that it would be beneficial if medical professionals provided support and allocated some time to help patients understand the AI technologies being used in clinical practice. Hence, understanding the fundamentals underpinning AI technology helps foster confidence among patients and increases their appreciation for the support provided by the technology:

A health professional who can also help me and guide me if there's something that I don't understand, or I'm missing a piece of this puzzle. So, a coach and educator. Yeah, someone who's got my back with the AI as well. So again, I just think we can't lose sight of that human touch and how we learn and digest and understand information. It's not just a transaction. [ID 9]

An intuitive, interactive AI app or technology was also mentioned as an important element of confidence. When patients use technology as part of their care, they want to ensure that their concerns, thoughts, and opinions are heard. When their

care provider was not physically present, patients expressed the desire for a connection. That is, despite the lack of a physical presence, patients preferred using a technology with interactive features to respond to their questions or concerns.

Theme 3: Establishing Data Governance and Validation of AI Technologies

Responsibility of Data Stewards

Participants expressed privacy concerns, such as how their health data would be used and shared, and for what purposes. In particular, participants mentioned fear of their personal data in apps being sold to private companies or used for illicit purposes:

I have strong objections to it being sold. I know the [organization] was making their data available to a private company at one point. And I know there are doctors in Ontario who feel that the health record is theirs, and they own it. And so, the information and it may be mine, but since they own the program that holds my data, they feel they have every right to sell it, and they do. So, I want more control over who gets to use it and why. And I mean, I think a lot of people would say, I'm fine for the public good, I'm fine with research that will benefit me, and people like me. But they draw the line at people making money from their personal data. [ID 10]

Participants voiced several concerns about the privacy of their health data and its potential for long-term use when entering web-based portals or apps. Many participants suggested the importance of choice regarding the types of information used for secondary purposes. They also expressed value in having the option to accept or reject the use of their information by third parties and to be able to remove their data, if desired. One participant worried about long-term consequences, such as familial genetic records being attached to future generations and potential lifetime implications from youth sharing personal information on mental health chatbots. Another felt it was important to understand how their health data were used to augment AI and its financial implications. Patients also wanted to be informed of how their health data would be protected, how to access their own data, who had access to it, and potential long-term consequences. Gatekeepers were identified as critical in ensuring the compliance and security of patient data as well as managing any regulatory risks:

I would like to know if there's any third parties going to see it. My other concern...Say the insurance, I tested positive for breast cancer, and it was a genetic one, I'm going through that right now. What how having AI and data out there on a computer without being shared with insurance companies, which is more likely to happen than it is right now. So, yeah, I would want to know how my privacy is being respected. And any third parties involved and any changes I'd want to be updated and if there were changes and third parties were going to see it, I'd have the choice of letting them or completely removing all my information. [ID 2]

Informed consent to access data, disclosure of use, and potential risks were stated as critical measures to protect patient privacy. Data protection and security were emphasized as key mitigation steps to ensure that patient data would not be disclosed. If data were shared without consent or accidentally, participants expressed the need for legal barriers, so that third-party companies would have no recourse. Participants desired apps to be verified by trusted sources, such as hospitals and the government, with transparency on the backend technologies deployed within them and how their data would be handled.

Quality Assurance and Validation of AI Technologies

Interestingly, participants also highlighted the need to understand more about how health care systems benefit from investment in AI technologies. They reported that this would help deliver care more effectively through the use of preventive tools and by identifying optimal treatment options. Some participants argued that AI technologies could contribute to additional health expenditures and further amplify the pressure on an overburdened health care system. In a public health system, it is essential to maximize benefits across the system and reduce costs.

Moreover, participants reported the need for governance and oversight in terms of quality of assurance and accessibility of technology. Participants emphasized that there should be a governing body that evaluates the technologies used in clinical care before endorsing them:

Then that becomes no different if there's no oversight or no background or no warnings about them or disclaimers, then it becomes just the same as people Googling everything. So, I would want it to be a better tool and a somewhat regulated tool or something so that it's actually endorsed by the medical community before it's available, or at least obviously they're not going to be able to control everything that's available on the Internet. But at least there would be some education to the public that to use the tools that we endorse or use the tool endorsed by your hospital or your province or whatever, there would be some kind of oversight. That's all I'm concerned about, that it just becomes the next version of Google. [ID7]

Participants preferred a regulated technology that was validated by the medical community before being available to the public. One participant mentioned that, without regulation, random apps would be produced and sold to hospitals.

Ensuring AI Technologies Used in Clinical Contexts Are Equitable and Inclusive

Participants would like to understand how AI technologies will be used in their health care, who would be using them, and for what reasons. One of the participants also mentioned that AI could be an unbiased solution for physicians to use in their care:

Oh, yeah, definitely as a tool to assist physicians, I think it would be great. And I think that there are circumstances where the artificial intelligence tool might do a better job than the doctor. Because you know, a lot of people in health care are...people have

preconceived notions about them, right. For instance, if somebody decides that you're a hysterical woman, you won't get the same care as you would if you had didn't have that notation in your health record. And so, I think that with the use of artificial intelligence, it takes out some of the bias. [ID 10]

Some participants reported the use of biased data for model development and the lack of diversity represented in data sets as problematic. Inherent biases are sometimes created when data sets are not heterogeneous, which can exclude vulnerable populations. Sex and racial disparities, for instance, can also be created if inherently biased data are included in data sets and applications:

I guess it really depends on who has actually set up the AI and what biases they have and what has actually been programmed into the system and if that's actually missing data, just because of the bias and missing marginalized populations or people that don't have a lot of money or are of a different race. And look, I just think there was something that I saw a while back about an app, you know, telling somebody had heart attack symptoms, and if it was male, it would say you should go to the hospital. But if it was female, it was like, oh, you don't have a heart attack. You have I'm guessing this was a while ago, I'm guessing probably anxiety! So, there's like sex differences, too. And so, I just wonder about the disparities that could be created, if it hasn't been created with the people that it's looking at. [ID 5]

The participants stressed the importance of ensuring that the training and testing data sets are heterogeneous and representative of the target population. Acknowledging these biases enables clinicians to make informed decisions and prevent any unintended consequences of patient care.

Discussion

Principal Findings

As new technologies and AI solutions emerge within health care, it is crucial to ensure that patients are included in the delivery of their own care. Advancements in digital technologies have revolutionized the possibilities of delivering optimal and patient-centric care in this continuously evolving health care ecosystem. Despite the rapid penetration of innovative technologies in clinical care, little is known about the effectiveness of AI technologies. The efficacy and long-term adoption of these technologies depend greatly on patient engagement and adherence [15]. McMahon [42] contended that patient engagement as part of medical education and continuing professional development is crucial in providing an opportunity for HCPs to develop their patient-centric skills, increase sensitivity to patient needs and values, and foster interprofessional collaborative practice. Patient expertise is based on their unique experiences of receiving care and the impact of the social determinants of health. Therefore, it is important to acknowledge and appreciate the value of these diverse patient viewpoints [43,44]. In addition, patient participation is reported to improve care providers'

communication skills and empathy and increase their awareness of patients' needs in marginalized communities [45-47].

This study aimed to understand patients' perspectives on how to better foster patient engagement in the uptake of AI technologies and what competencies they believe are essential in preparing HCPs for digital care. Through semistructured interviews with patient partners, three predominant themes emerged: (1) cultivating patients' trust, (2) fostering patient engagement, and (3) establishing data governance and validation of AI technologies. Participants in both urban and rural settings highlighted similar ideas with regard to AI adoption.

In a recent scoping review, Charow et al [34] identified key competencies that are currently taught as part of the AI curriculum and what programs should be taught. The authors used Bloom Learning Taxonomy to group curriculum topics [34]. Table 3 illustrates the overlap of competencies identified in the scoping review and highlighted by the participants in this study.

As technologies are being integrated within care settings, participants in this study emphasized that it is important for HCPs to acknowledge how data are acquired and processed and explain a rationale when making decisions. Interestingly, the psychomotor and affective domains of Bloom Learning Taxonomy were reiterated by participants. Critical appraisal, ethical and legal considerations, communication, interpersonal skills, empathy, compassion, and emotional responsiveness were highlighted as important competencies to minimize the negative implications of AI integration at the point of care.

This study highlights the importance of establishing trust and transparency as part of the patient-clinician relationship. Many participants stated that lack of transparency in data access and use could potentially erode their trust in using AI for care delivery. This was in line with a previous study [30], which suggested that physicians must have a thorough knowledge of the AI technologies used and be prepared to provide a coherent rationale when making clinical decisions. For instance, if a patient is diagnosed with cancer, they would want to understand how AI technology arrived at that decision [48]. What becomes a challenge, however, is that advanced AI technologies are often built using complex algorithms, which may be difficult to explain, even if clinicians have the technical expertise [48]. In a qualitative study that examined patient privacy perspectives on health information exchange, trust was identified as a key antecedent for establishing effective patient-clinician relationships [49]. Transparent communication regarding the use of AI technologies serves as an initial step toward cultivating trust [49]. The authors noted a significant association between patients' trust in clinicians and their willingness to share personal health information [49].

Patients believe the clinician's presence is important, particularly when discussing sensitive information regarding their care. AI technologies should support existing patient care and not replace physician interactions. Similar to our study, previous research indicated that patients valued the interaction with the clinician rather than with AI technology alone [29]. AI technologies can potentially diminish clinician-patient interactions and jeopardize the humanistic facet of patient care [15,50]. Patients who

interacted only with AI technologies in their care reported a lack of compassion and empathy [19,21,22] and a limited opportunity for patients to ask follow-up questions, discuss treatment options, and receive emotional support [19,21]. Davenport and Kalakota [48] further reinforced this point, highlighting the importance of establishing an empathetic relationship between clinicians and patients. In other studies, patients specified that the AI output should be verified by the physician for accuracy [22] and be used as a second opinion to inform clinical decisions [9,19]. In the event of a disagreement between the physician and the AI technology, patients favor the physician's judgment as the final decision [9,22]. Yang et al [21] reported that AI can serve as a copilot in automating tasks and optimizing the quality of care. More importantly, the literature emphasizes the role of providers in decision-making, as they need to adapt the AI results based on the uniqueness of each patient and their circumstances [9].

Engaging patients in proactive care leads to better patient experience and improved health system outcomes [48]. The findings from this study suggest that education on AI innovations helps to create awareness and foster confidence among patients. As a result, patients' self-efficacy increases, enabling them to be knowledgeable and competent in safely navigating a digitized health care environment. This also contributes to the increased acceptance of AI technologies in practical settings to enhance the quality of care. Recent studies on patient perspectives on the use of AI in health care reported that it is critical for patients to be educated on the threats of AI technologies in an ever-increasing technology-enabled care environment [50,51]. Cultivating a strong culture of cybervigilance across this new digital space is vital for delivering care and ensuring that large amounts of sensitive and valuable data in vulnerable systems are protected. Moreover, Kovarik [52] reported that patients should be educated on the fundamentals of AI, which will be valuable when discussing diagnoses and treatment options.

Furthermore, the findings of this study underline the need for data stewards and regulations to ensure the protection and confidentiality of patient data. Consistent with previous literature, patients reported high levels of concern toward the misuse of their personal health information [15,48,51]. Patients in this study also expressed privacy concerns, such as how their health data would be used, how their data would be shared, and for what purposes. This ambivalence has resulted in increased fear among patients, and the need for choice and autonomy. Participants stated that it was important to have a choice in terms of consenting to what information they would prefer to opt-in or opt-out for secondary use of data. In a review article on the practical implementation of AI technologies, the authors asserted that cybersecurity measures need to be implemented to address concerns about the inappropriate use of patient data

[53]. A few studies have reported that patients feared that their personal health information might be not anonymized or be used for profit by insurance and third-party companies [15,50]. In one study, patients perceived that insurance companies could use AI technologies to discern new information about their health and make changes to their premiums [9].

Oversight and regulatory measures are necessary to ensure the confidentiality of patient data and to protect against nefarious acts [9]. The AI implementation toolkit developed by Canada Health Infoway provides guidance on an AI governance framework [54]. This framework consists of 3 key constructs that oversee the responsible and ethical implementation of AI technologies: people, policies, and procedures [54]. The people construct consists of skillsets required to form a committee that provides procedural and practical guidance for AI implementation [54]. Policies focus on providing directions for risk considerations related to AI [54]. Procedures provide operational guidance on implementation aspects, including risk assessment, data testing, and monitoring [54]. Establishing governance structures is pivotal in monitoring ethical issues and mitigating any negative repercussions as a result of AI implementation in a milieu of increasing vulnerability to data breaches [48]. Matheny et al [3] delineated that it is imperative to involve patients and their families when developing regulatory and legislative solutions regarding the use of AI technologies in clinical contexts.

Finally, the participants noted the importance of examining implicit biases to ensure that AI technologies are inclusive and equitable. Biases in data sets may pose challenges in generalizing results and further exacerbate health inequities as well as discriminatory practices. This point was reinforced in a nominal group technique study that emphasized the negative implications of using homogenous data sets for developing algorithms [23]. One example of this is when AI models are developed based on data from a single health care institution, which may not be representative of a larger population [55]. The literature also reports that developers could inadvertently integrate their biases into the model development process [9]. Daneshjou et al [56] noted that there are no standards for describing data sets used for AI model development. Descriptions of data sets could aid in a better understanding of models and any underlying biases. Interestingly, our study also accentuated the notion of using AI technologies to reduce bias from a patient perspective. In health care, clinicians sometimes have preconceived notions about their patients; hence, a patient may not receive the same care as they would if they did not have that notation in their health records. Participants believed that AI technologies could remove some of the preconceived ideas and perceptions that contribute to the marginalization of specific populations when providing care, thus creating a more equitable and inclusive care environment.

Table 3. Overlap of competencies identified in the scoping review and highlighted by participants in this study.

Bloom taxonomy domain	Competencies identified in the scoping review (Charow et al [34])	Competencies highlighted by participants in this study and the scoping review (Charow et al [34])
Cognitive	<ul style="list-style-type: none"> • Fundamentals of AI^a • Implementation of AI • Big data • Data science, machine learning, and statistics • Multidisciplinary collaboration • Strengths and limitations of AI • Predictive analytics • Economic considerations • EHR^b fundamentals 	<ul style="list-style-type: none"> • Ethics and legal issues • Data governance
Psychomotor	<ul style="list-style-type: none"> • Analytical • Problem solving • Product development • Data visualization 	<ul style="list-style-type: none"> • Interpretation • Communication • Critical appraisal • Medical decision-making • Cultivation of compassion and empathy
Affective	<ul style="list-style-type: none"> • Change management • Adoption of AI 	<ul style="list-style-type: none"> • Perceptions of humanistic AI-enabled care • Create and sustain a culture of trust and transparency with stakeholders and patients

^aAI: artificial intelligence.

^bEHR: electronic health record.

Limitations

The findings of this study should be examined in light of these limitations. A limitation of this study is that the study population included no individuals in the age range of 0 to 40 years. Despite the less frequent use of health care services in this age group, they may represent a more technology-savvy population. This study provides diverse perspectives from rural and urban settings in Canada, as context plays a pivotal role in influencing the uptake of technology. This study provides a nuanced understanding of patient perceptions in both settings and how their perceptions may be similar. The interviews were conducted until theoretical saturation was achieved (n=12). In addition, a rigorous analytical approach was adopted, including iterative discussions with the research team and patient partners to validate emerging themes. Another limitation of this study was the recruitment of predominantly female patients, contributing to an underrepresentation of male voices. Demographic data such as race, ethnicity, employment, disability, and language were not collected, as the purposive sampling attempted to recruit participants based on comfort with the technology and the contexts in which they received care.

Conclusions

This study revealed that to successfully adopt AI technologies in care settings, it is crucial to foster patient trust, build continued partnerships with patients, and establish data governance and validation of AI technologies. As we shift to a digital form of care, AI innovations are being rapidly adopted and implemented within the clinical ecosystem at a fast pace to advance the delivery of patient care and enhance efficiency at a systems level. Rather than AI becoming a replacement for humanistic care, AI and care providers play a synergetic role in the future of digital care. Understanding the needs and values of patients helps ensure the safe, effective, and responsible use of AI. Patient engagement helps to provide a real-world perspective and coconstruct knowledge from an end-user standpoint, thus ensuring that AI innovations are successfully integrated into practice settings. The findings of this study have implications for all stakeholders with accountability to ensure that patients are actively engaged in sustaining safe and high-quality care.

Acknowledgments

Accelerating the appropriate adoption of AI in health care by building new knowledge, skills, and capacities in the Canadian health care professions is funded by the Government of Canada's Future Skills Centre.

Accélérer l'adoption appropriée de l'intelligence artificielle dans la santé en développant de nouvelles connaissances, compétences et capacités pour les professionnels de santé canadiens est financé par le Centre des Compétences futures du gouvernement du Canada.

The authors wish to thank all the participants for their time and contribution to the study. They also thank Ms Sarmini Balakumar for her support and assistance with data analysis.

Authors' Contributions

DW conceived the study and revised all drafts. Each semistructured interview was conducted by 4 members of the research team (MC, MZ, SY, and TJ). MC, MZ, SY, and TJ coded the interview transcripts and inductively analyzed the data. TJ and SY prepared the initial manuscript draft. All the authors have read and approved the final manuscript.

Conflicts of Interest

None declared.

Multimedia Appendix 1

Patient context.

[\[DOCX File, 17 KB - ai_v2i1e40973_app1.docx\]](#)

References

1. Tang A, Tam R, Cadrin-Chênevert A, Guest W, Chong J, Barfett J, Canadian Association of Radiologists (CAR) Artificial Intelligence Working Group. Canadian Association of Radiologists white paper on artificial intelligence in radiology. *Can Assoc Radiol J* 2018 May;69(2):120-135 [FREE Full text] [doi: [10.1016/j.carj.2018.02.002](https://doi.org/10.1016/j.carj.2018.02.002)] [Medline: [29655580](https://pubmed.ncbi.nlm.nih.gov/29655580/)]
2. Simon M, Baur C, Guastello S, Ramiah K, Tuft J, Wisdom K, et al. Patient and family engaged care: an essential element of health equity. *NAM Perspect* 2020 Jul 13;2020:1-26 [FREE Full text] [doi: [10.31478/202007a](https://doi.org/10.31478/202007a)] [Medline: [35291751](https://pubmed.ncbi.nlm.nih.gov/35291751/)]
3. Matheny ME, Whicher D, Israni ST. Artificial intelligence in health care: a report from the National Academy of Medicine. *JAMA* 2020 Feb 11;323(6):509-510 [FREE Full text] [doi: [10.1001/jama.2019.21579](https://doi.org/10.1001/jama.2019.21579)] [Medline: [31845963](https://pubmed.ncbi.nlm.nih.gov/31845963/)]
4. Strome E. Building a learning health system for Canadians. Canadian Institute for Advanced Research. Toronto, Canada: Canadian Institute for Advanced Research; 2020 Jul. URL: <https://cifar.ca/wp-content/uploads/2020/11/AI4Health-report-ENG-10-F.pdf> [accessed 2022-01-05]
5. Banerjee S, Alsop P, Jones L, Cardinal RN. Patient and public involvement to build trust in artificial intelligence: a framework, tools, and case studies. *Patterns (N Y)* 2022 Jun 10;3(6):100506 [FREE Full text] [doi: [10.1016/j.patter.2022.100506](https://doi.org/10.1016/j.patter.2022.100506)] [Medline: [35755870](https://pubmed.ncbi.nlm.nih.gov/35755870/)]
6. Donia J, Shaw JA. Co-design and ethical artificial intelligence for health: an agenda for critical research and practice. *Big Data Soc* 2021 Dec 17;8(2):205395172110652 [FREE Full text] [doi: [10.1177/20539517211065248](https://doi.org/10.1177/20539517211065248)]
7. Zidaru T, Morrow EM, Stockley R. Ensuring patient and public involvement in the transition to AI-assisted mental health care: a systematic scoping review and agenda for design justice. *Health Expect* 2021 Aug;24(4):1072-1124 [FREE Full text] [doi: [10.1111/hex.13299](https://doi.org/10.1111/hex.13299)] [Medline: [34118185](https://pubmed.ncbi.nlm.nih.gov/34118185/)]
8. Richardson JP, Curtis S, Smith C, Pacyna J, Zhu X, Barry B, et al. A framework for examining patient attitudes regarding applications of artificial intelligence in healthcare. *Digit Health* 2022 Mar 24;8:20552076221089084 [FREE Full text] [doi: [10.1177/20552076221089084](https://doi.org/10.1177/20552076221089084)] [Medline: [35355806](https://pubmed.ncbi.nlm.nih.gov/35355806/)]
9. Richardson JP, Smith C, Curtis S, Watson S, Zhu X, Barry B, et al. Patient apprehensions about the use of artificial intelligence in healthcare. *NPJ Digit Med* 2021 Sep 21;4(1):140 [FREE Full text] [doi: [10.1038/s41746-021-00509-1](https://doi.org/10.1038/s41746-021-00509-1)] [Medline: [34548621](https://pubmed.ncbi.nlm.nih.gov/34548621/)]
10. Higgins T, Larson E, Schnell R. Unraveling the meaning of patient engagement: a concept analysis. *Patient Educ Couns* 2017 Jan;100(1):30-36. [doi: [10.1016/j.pec.2016.09.002](https://doi.org/10.1016/j.pec.2016.09.002)] [Medline: [27665500](https://pubmed.ncbi.nlm.nih.gov/27665500/)]
11. Patient engagement. Canadian Institutes of Health Research. 2019 May 27. URL: <https://cihr-irsc.gc.ca/e/45851.html> [accessed 2022-01-05]
12. Study: quebecers open to AI in healthcare. Canadian Healthcare Technology. 2022 Jan 26. URL: <https://www.canhealth.com/2022/01/26/study-quebecers-open-to-ai-in-healthcare/> [accessed 2022-02-05]
13. Adams SJ, Tang R, Babyn P. Patient perspectives and priorities regarding artificial intelligence in radiology: opportunities for patient-centered radiology. *J Am Coll Radiol* 2020 Aug;17(8):1034-1036. [doi: [10.1016/j.jacr.2020.01.007](https://doi.org/10.1016/j.jacr.2020.01.007)] [Medline: [32068006](https://pubmed.ncbi.nlm.nih.gov/32068006/)]
14. Dieng M, Smit AK, Hersch J, Morton RL, Cust AE, Irwig L, et al. Patients' views about skin self-examination after treatment for localized melanoma. *JAMA Dermatol* 2019 Aug 01;155(8):914-921 [FREE Full text] [doi: [10.1001/jamadermatol.2019.0434](https://doi.org/10.1001/jamadermatol.2019.0434)] [Medline: [31090868](https://pubmed.ncbi.nlm.nih.gov/31090868/)]
15. Jutzi TB, Krieghoff-Henning EI, Holland-Letz T, Utikal JS, Hauschild A, Schandendorf D, et al. Artificial intelligence in skin cancer diagnostics: the patients' perspective. *Front Med (Lausanne)* 2020 Jun 02;7:233 [FREE Full text] [doi: [10.3389/fmed.2020.00233](https://doi.org/10.3389/fmed.2020.00233)] [Medline: [32671078](https://pubmed.ncbi.nlm.nih.gov/32671078/)]
16. McCradden MD, Baba A, Saha A, Ahmad S, Boparai K, Fadaiefard P, et al. Ethical concerns around use of artificial intelligence in health care research from the perspective of patients with meningioma, caregivers and health care providers: a qualitative study. *CMAJ Open* 2020 Feb 18;8(1):E90-E95 [FREE Full text] [doi: [10.9778/cmajo.20190151](https://doi.org/10.9778/cmajo.20190151)] [Medline: [32071143](https://pubmed.ncbi.nlm.nih.gov/32071143/)]

17. 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 Oct 28;10(10):e039798 [[FREE Full text](#)] [doi: [10.1136/bmjopen-2020-039798](https://doi.org/10.1136/bmjopen-2020-039798)] [Medline: [33115901](#)]
18. Meyer AN, Giardina TD, Spitzmueller C, Shahid U, Scott TM, Singh H. Patient perspectives on the usefulness of an artificial intelligence-assisted symptom checker: cross-sectional survey study. *J Med Internet Res* 2020 Jan 30;22(1):e14679 [[FREE Full text](#)] [doi: [10.2196/14679](https://doi.org/10.2196/14679)] [Medline: [32012052](#)]
19. Nelson CA, Pérez-Chada LM, Creadore A, Li SJ, Lo K, Manjaly P, et al. Patient perspectives on the use of artificial intelligence for skin cancer screening: a qualitative study. *JAMA Dermatol* 2020 May 01;156(5):501-512 [[FREE Full text](#)] [doi: [10.1001/jamadermatol.2019.5014](https://doi.org/10.1001/jamadermatol.2019.5014)] [Medline: [32159733](#)]
20. Ongena YP, Haan M, Yakar D, Kwee TC. Patients' views on the implementation of artificial intelligence in radiology: development and validation of a standardized questionnaire. *Eur Radiol* 2020 Feb;30(2):1033-1040 [[FREE Full text](#)] [doi: [10.1007/s00330-019-06486-0](https://doi.org/10.1007/s00330-019-06486-0)] [Medline: [31705254](#)]
21. Yang L, Ene IC, Belaghi RA, Koff D, Stein N, Santaguida PL. Stakeholders' perspectives on the future of artificial intelligence in radiology: a scoping review. *Eur Radiol* 2022 Mar;32(3):1477-1495. [doi: [10.1007/s00330-021-08214-z](https://doi.org/10.1007/s00330-021-08214-z)] [Medline: [34545445](#)]
22. Lennartz S, Dratsch T, Zopfs D, Persigehl T, Maintz D, Große Hokamp N, et al. Use and control of artificial intelligence in patients across the medical workflow: single-center questionnaire study of patient perspectives. *J Med Internet Res* 2021 Feb 17;23(2):e24221 [[FREE Full text](#)] [doi: [10.2196/24221](https://doi.org/10.2196/24221)] [Medline: [33595451](#)]
23. Musbahi O, Syed L, Le Feuvre P, Cobb J, Jones G. Public patient views of artificial intelligence in healthcare: a nominal group technique study. *Digit Health* 2021 Dec 15;7:20552076211063682 [[FREE Full text](#)] [doi: [10.1177/20552076211063682](https://doi.org/10.1177/20552076211063682)] [Medline: [34950499](#)]
24. Young AT, Amara D, Bhattacharya A, Wei ML. Patient and general public attitudes towards clinical artificial intelligence: a mixed methods systematic review. *Lancet Digit Health* 2021 Sep;3(9):e599-e611 [[FREE Full text](#)] [doi: [10.1016/S2589-7500\(21\)00132-1](https://doi.org/10.1016/S2589-7500(21)00132-1)] [Medline: [34446266](#)]
25. Balthazar P, Harri P, Prater A, Safdar NM. Protecting your patients' interests in the era of big data, artificial intelligence, and predictive analytics. *J Am Coll Radiol* 2018 Mar;15(3 Pt B):580-586 [[FREE Full text](#)] [doi: [10.1016/j.jacr.2017.11.035](https://doi.org/10.1016/j.jacr.2017.11.035)] [Medline: [29402532](#)]
26. Panch T, Duralde E, Mattie H, Kotecha G, Celi LA, Wright M, et al. A distributed approach to the regulation of clinical AI. *PLOS Digit Health* 2022 May 26;1(5):e0000040 [[FREE Full text](#)] [doi: [10.1371/journal.pdig.0000040](https://doi.org/10.1371/journal.pdig.0000040)]
27. Esmailzadeh P, Mirzaei T, Dharanikota S. Patients' perceptions toward human-artificial intelligence interaction in health care: experimental study. *J Med Internet Res* 2021 Nov 25;23(11):e25856 [[FREE Full text](#)] [doi: [10.2196/25856](https://doi.org/10.2196/25856)] [Medline: [34842535](#)]
28. 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 [[FREE Full text](#)] [doi: [10.1016/j.ibmed.2020.100001](https://doi.org/10.1016/j.ibmed.2020.100001)]
29. 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: HICSS '21; January 5-8, 2021; Kauai, HI, USA p. 5463-5472 URL: <https://scholarspace.manoa.hawaii.edu/handle/10125/71284> [doi: [10.24251/HICSS.2021.664](https://doi.org/10.24251/HICSS.2021.664)]
30. Nundy S, Montgomery T, Wachter RM. Promoting trust between patients and physicians in the era of artificial intelligence. *JAMA* 2019 Aug 13;322(6):497-498 [[FREE Full text](#)] [doi: [10.1001/jama.2018.20563](https://doi.org/10.1001/jama.2018.20563)] [Medline: [31305873](#)]
31. 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 May;25(2):45-46 [[FREE Full text](#)] [doi: [10.1136/ebmental-2022-300489](https://doi.org/10.1136/ebmental-2022-300489)] [Medline: [35444002](#)]
32. Kerasidou A. Artificial intelligence and the ongoing need for empathy, compassion and trust in healthcare. *Bull World Health Organ* 2020 Apr 01;98(4):245-250 [[FREE Full text](#)] [doi: [10.2471/BLT.19.237198](https://doi.org/10.2471/BLT.19.237198)] [Medline: [32284647](#)]
33. Clavel N, Paquette J, Dumez V, Del Grande C, Ghadiri DP, Pomey MP, et al. Patient engagement in care: a scoping review of recently validated tools assessing patients' and healthcare professionals' preferences and experience. *Health Expect* 2021 Dec;24(6):1924-1935 [[FREE Full text](#)] [doi: [10.1111/hex.13344](https://doi.org/10.1111/hex.13344)] [Medline: [34399008](#)]
34. Charow R, Jeyakumar T, Younus S, Dolatabadi E, Sahlia M, Al-Mouaswas D, et al. Artificial intelligence education programs for health care professionals: scoping review. *JMIR Med Educ* 2021 Dec 13;7(4):e31043 [[FREE Full text](#)] [doi: [10.2196/31043](https://doi.org/10.2196/31043)] [Medline: [34898458](#)]
35. Campbell S, Greenwood M, Prior S, Shearer T, Walkem K, Young S, et al. Purposive sampling: complex or simple? Research case examples. *J Res Nurs* 2020 Dec;25(8):652-661 [[FREE Full text](#)] [doi: [10.1177/1744987120927206](https://doi.org/10.1177/1744987120927206)] [Medline: [34394687](#)]
36. Robinson RS. Purposive sampling. In: Michalos AC, editor. *Encyclopedia of Quality of Life and Well-Being Research*. Dordrecht, The Netherlands: Springer; 2014:5243-5245.
37. Johnson TP. Snowball sampling: introduction. In: Armitage P, Colton T, editors. *Encyclopedia of Biostatistics*. Hoboken, NJ, USA: John Wiley & Sons; Sep 29, 2014.

38. Dodgson JE. Reflexivity in qualitative research. *J Hum Lact* 2019 May;35(2):220-222 [[FREE Full text](#)] [doi: [10.1177/0890334419830990](https://doi.org/10.1177/0890334419830990)] [Medline: [30849272](#)]
39. Creswell JW, Poth CN. *Qualitative Inquiry and Research Design: Choosing Among Five Approaches*. 4th edition. Thousand Oaks, CA, USA: Sage Publications; 2017.
40. Braun V, Clarke V. Using thematic analysis in psychology. *Qual Res Psychol* 2008 Jul 21;3(2):77-101 [[FREE Full text](#)] [doi: [10.1191/1478088706qp063oa](https://doi.org/10.1191/1478088706qp063oa)]
41. Braun V, Clarke V. One size fits all? What counts as quality practice in (reflexive) thematic analysis? *Qual Res Psychol* 2020 Aug 12;18(3):328-352 [[FREE Full text](#)] [doi: [10.1080/14780887.2020.1769238](https://doi.org/10.1080/14780887.2020.1769238)]
42. McMahon GT. Learning together: engaging patients as partners in CPD. *J Contin Educ Health Prof* 2021 Oct 01;41(4):268-272 [[FREE Full text](#)] [doi: [10.1097/CEH.0000000000000388](https://doi.org/10.1097/CEH.0000000000000388)] [Medline: [34609358](#)]
43. Hill G, Thompson G, Willis S, Hodgson D. Embracing service user involvement in radiotherapy education: a discussion paper. *Radiography* 2014 Feb;20(1):82-86 [[FREE Full text](#)] [doi: [10.1016/j.radi.2013.08.007](https://doi.org/10.1016/j.radi.2013.08.007)]
44. Szumacher E. Patients' engagement in medical education. *J Cancer Educ* 2019 Apr;34(2):203-204 [[FREE Full text](#)] [doi: [10.1007/s13187-019-01496-4](https://doi.org/10.1007/s13187-019-01496-4)] [Medline: [30852788](#)]
45. Dijk SW, Duijzer EJ, Wienold M. Role of active patient involvement in undergraduate medical education: a systematic review. *BMJ Open* 2020 Jul 27;10(7):e037217 [[FREE Full text](#)] [doi: [10.1136/bmjopen-2020-037217](https://doi.org/10.1136/bmjopen-2020-037217)] [Medline: [32718925](#)]
46. Henry-Noel N, Paton M, Wong R, Dawdy K, Karim A, Soliman H, et al. Patient engagement in the continuing professional development programs within the department of radiation oncology at the University of Toronto (UTDRO): a qualitative study. *J Med Imaging Radiat Sci* 2022 Jun;53(2):256-263 [[FREE Full text](#)] [doi: [10.1016/j.jmir.2022.03.003](https://doi.org/10.1016/j.jmir.2022.03.003)] [Medline: [35393257](#)]
47. Towle A, Bainbridge L, Godolphin W, Katz A, Kline C, Lown B, et al. Active patient involvement in the education of health professionals. *Med Educ* 2010 Jan;44(1):64-74 [[FREE Full text](#)] [doi: [10.1111/j.1365-2923.2009.03530.x](https://doi.org/10.1111/j.1365-2923.2009.03530.x)] [Medline: [20078757](#)]
48. Davenport T, Kalakota R. The potential for artificial intelligence in healthcare. *Future Healthc J* 2019 Jun;6(2):94-98 [[FREE Full text](#)] [doi: [10.7861/futurehosp.6-2-94](https://doi.org/10.7861/futurehosp.6-2-94)] [Medline: [31363513](#)]
49. Shen N, Sequeira L, Silver MP, Carter-Langford A, Strauss J, Wiljer D. Patient privacy perspectives on health information exchange in a mental health context: qualitative study. *JMIR Ment Health* 2019 Nov 13;6(11):e13306 [[FREE Full text](#)] [doi: [10.2196/13306](https://doi.org/10.2196/13306)] [Medline: [31719029](#)]
50. 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 Jun 14;2:53 [[FREE Full text](#)] [doi: [10.1038/s41746-019-0132-y](https://doi.org/10.1038/s41746-019-0132-y)] [Medline: [31304399](#)]
51. Aggarwal R, Farag S, Martin G, Ashrafian H, Darzi A. Patient perceptions on data sharing and applying artificial intelligence to health care data: cross-sectional survey. *J Med Internet Res* 2021 Aug 26;23(8):e26162 [[FREE Full text](#)] [doi: [10.2196/26162](https://doi.org/10.2196/26162)] [Medline: [34236994](#)]
52. Kovarik CL. Patient perspectives on the use of artificial intelligence. *JAMA Dermatol* 2020 May 01;156(5):493-494 [[FREE Full text](#)] [doi: [10.1001/jamadermatol.2019.5013](https://doi.org/10.1001/jamadermatol.2019.5013)] [Medline: [32159724](#)]
53. He J, Baxter SL, Xu J, Xu J, Zhou X, Zhang K. The practical implementation of artificial intelligence technologies in medicine. *Nat Med* 2019 Jan;25(1):30-36 [[FREE Full text](#)] [doi: [10.1038/s41591-018-0307-0](https://doi.org/10.1038/s41591-018-0307-0)] [Medline: [30617336](#)]
54. Toolkit for Implementers of Artificial Intelligence in Health Care. Canada Health Infoway. 2021 Dec. URL: <https://ittechreports.com/toolkit-for-implementers-of-artificial-intelligence-in-health-care/> [accessed 2022-04-22]
55. Yu KH, Beam AL, Kohane IS. Artificial intelligence in healthcare. *Nat Biomed Eng* 2018 Oct;2(10):719-731 [[FREE Full text](#)] [doi: [10.1038/s41551-018-0305-z](https://doi.org/10.1038/s41551-018-0305-z)] [Medline: [31015651](#)]
56. 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. [doi: [10.1001/jamadermatol.2021.3129](https://doi.org/10.1001/jamadermatol.2021.3129)] [Medline: [34550305](#)]

Abbreviations

AI: artificial intelligence

HCP: health care professional

Edited by K El Emam, B Malin; submitted 11.07.22; peer-reviewed by Z Azizi, D Radhakrishnan, C Lai; comments to author 04.10.22; revised version received 29.11.22; accepted 29.12.22; published 02.03.23.

Please cite as:

Jeyakumar T, Younus S, Zhang M, Clare M, Charow R, Karsan I, Dhalla A, Al-Mouaswas D, Scandiffio J, Aling J, Salhia M, Lalani N, Overholt S, Wiljer D

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 AI 2023;2:e40973

URL: <https://ai.jmir.org/2023/1/e40973>

doi: [10.2196/40973](https://doi.org/10.2196/40973)

PMID: [38875561](https://pubmed.ncbi.nlm.nih.gov/38875561/)

©Tharshini Jeyakumar, Sarah Younus, Melody Zhang, Megan Clare, Rebecca Charow, Inaara Karsan, Azra Dhalla, Dalia Al-Mouaswas, Jillian Scandiffio, Justin Aling, Mohammad Salhia, Nadim Lalani, Scott Overholt, David Wiljer. Originally published in JMIR AI (<https://ai.jmir.org>), 02.03.2023. This is an open-access article distributed under the terms of the Creative Commons Attribution License (<https://creativecommons.org/licenses/by/4.0/>), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR AI, is properly cited. The complete bibliographic information, a link to the original publication on <https://www.ai.jmir.org/>, as well as this copyright and license information must be included.

Original Paper

Artificial Intelligence in Health Care—Understanding Patient Information Needs and Designing Comprehensible Transparency: Qualitative Study

Renee Robinson¹, MSPharm, MPH, MBA, PharmD; Cara Liday², PharmD; Sarah Lee³, BA, BSc; Ishan C Williams⁴, PhD; Melanie Wright³, PhD; Sungjoon An³; Elaine Nguyen³, MPH, PharmD

¹College of Pharmacy, Idaho State University, Anchorage, AK, United States

²College of Pharmacy, Idaho State University, Pocatello, ID, United States

³College of Pharmacy, Idaho State University, Meridian, ID, United States

⁴School of Nursing, University of Virginia, Charlottesville, VA, United States

Corresponding Author:

Elaine Nguyen, MPH, PharmD

College of Pharmacy

Idaho State University

1311 E Central Dr

Meridian, ID, 83642

United States

Phone: 1 208 373 1829

Fax: 1 208 373 1834

Email: elainenguyen@isu.edu

Abstract

Background: Artificial intelligence (AI) is a branch of computer science that uses advanced computational methods, such as machine learning (ML), to calculate and predict health outcomes and address patient and provider health needs. While these technologies show great promise for improving health care, especially in diabetes management, there are usability and safety concerns for both patients and providers about the use of AI/ML in health care management.

Objective: We aimed to support and ensure safe use of AI/ML technologies in health care; thus, the team worked to better understand (1) patient information and training needs, (2) the factors that influence patients' perceived value and trust in AI/ML health care applications, and (3) how best to support safe and appropriate use of AI/ML-enabled devices and applications among people living with diabetes.

Methods: To understand general patient perspectives and information needs related to the use of AI/ML in health care, we conducted a series of focus groups (n=9) and interviews (n=3) with patients (n=41) and interviews with providers (n=6) in Alaska, Idaho, and Virginia. Grounded theory guided data gathering, synthesis, and analysis. Thematic content and constant comparison analysis were used to identify relevant themes and subthemes. Inductive approaches were used to link data to key concepts, including preferred patient-provider interactions and patient perceptions of trust, accuracy, value, assurances, and information transparency.

Results: Key summary themes and recommendations focused on (1) patient preferences for AI/ML-enabled device and application information, (2) patient and provider AI/ML-related device and application training needs, (3) factors contributing to patient and provider trust in AI/ML-enabled devices and applications, and (4) AI/ML-related device and application functionality and safety considerations. A number of participants (patients and providers) made recommendations to improve device functionality to guide information and labeling mandates (eg, link to online video resources and provide access to 24/7 live in-person or virtual emergency support). Other patient recommendations included (1) providing access to practice devices, (2) providing connections to local supports and reputable community resources, and (3) simplifying the display and alert limits.

Conclusions: Recommendations from both patients and providers could be used by federal oversight agencies to improve utilization of AI/ML monitoring of technology use in diabetes, improving device safety and efficacy.

(JMIR AI 2023;2:e46487) doi:[10.2196/46487](https://doi.org/10.2196/46487)

KEYWORDS

artificial intelligence; machine learning; diabetes; equipment safety; equipment design; health care

Introduction

Artificial intelligence (AI), a branch of computer science, attempts to build devices and software programs that explore and gather new knowledge, learn, and apply reasoning [1,2]. Machine learning (ML), a term often used interchangeably with AI, differs from AI in that in ML computer systems are able to adapt without following explicit instructions, using algorithms and statistical models to analyze and draw inferences from patterns in data [3,4]. Research in non-health care fields suggests that accountability is the most important attribute of AI, with fairness, security, privacy, and accuracy rated to have similarly high importance, and that transparent and comprehensible AI/ML systems are preferred [5-7]. Among the few studies that have explored patient perceptions of AI and related digital health applications, accuracy of decisions and patient empowerment have been identified as the 2 most important criteria [5,6]. In fact, a recent survey of health care workers in India found that technical skills, ethical concerns, and risk mitigation strategies were 3 key factors influencing perceptions regarding AI/ML use and that AI has a strong positive impact on patient cognitive engagement with health technologies [8].

As use of AI/ML in the health care arena is rapidly expanding, greater than expected benefits and patient outcomes have been seen [1]. Examples of AI/ML applications include but are not limited to diagnostic supports, image interpretation, tools that support rapid or automated data capture, and disease management [1,2]. In fact, recent studies have explored use of AI/ML in primary care [9] to support clinical decision-making and treatment management decisions for a number of chronic conditions, such as cardiovascular disease [10], mental health [11], and diabetes care [2]. However, little is known about how patients and providers feel about use of AI/ML in chronic disease management, if unmet AI/ML training needs influence AI/ML adoption, and most importantly, how barriers should be addressed (eg, labeling, training, and required supports). Left unaddressed, AI/ML concerns (eg, potential interpretation errors and data privacy issues) and use in nonrepresentative samples (eg, educated, well-resourced populations), could contribute to lack of patient and provider trust in AI/ML applications, health inequities, reduced efficacy, and poor patient outcomes, as well as preventable safety concerns [7,12,13].

The US Food and Drug Administration (FDA) is responsible for protecting public health by ensuring the safety, effectiveness, quality, and security of drugs, biological products, medical devices, and software (eg, mobile health apps) [14]. In 2014, the FDA established the Patient Engagement Advisory Committee (PEAC) to ensure safe and effective AI/ML implementation in the health care setting. The PEAC, made up of patients and providers, is responsible for premarket review of AI/ML devices, guiding device labeling requirements, and supporting “transparency and real-world performance monitoring” to ensure safe and effective AI/ML use from premarket development through the postmarketing period

[14,15]. The primary objective of this qualitative inquiry is to build upon the work of the FDA and PEAC to (1) understand general patient AI/ML information needs, (2) understand factors that influence patients’ perceived valuing of and trust in AI/ML devices to support diabetes management, and (3) guide current and future FDA AI/ML labeling requirements to ensure the appropriate information is accessible and supports safe and effective use of AI/ML-enabled devices.

Methods

Overview

Barriers to technology utilization (eg, understanding, access, and perceived need) differ by population and geographic region (eg, access in rural, underresourced, and ethnically diverse communities) [16,17]. Patients (and providers) may have limited awareness of the many AI/ML applications available to support patient health management. Assumed AI/ML application complexity, novelty, and costs make it difficult for patients to recognize and communicate their reservations and management needs with providers (eg, their general perceptions of the value of the relevant technologies, unmet information needs, necessary regulatory concerns, and assurances required to trust AI/ML applications) [17-20]. Due to the variety and relative maturity of available AI/ML diabetes management and prevention applications, we chose to focus on perceptions, information, and implementation needs of patients and providers considering and using AI/ML applications to manage their diabetes.

Setting

To understand general patient perspectives and information needs related to the use of AI/ML in health care, we conducted a series of 9 focus groups and 3 interviews that included a total of 41 patients and interviews with 6 providers, including nurse case managers, pharmacists, physicians, and an endocrinologist serving 3 different patient populations in Alaska (n=9), Idaho (n=23), and Virginia (n=8). Within the context of this study, members of the research team and target research population were part of the community of interest (individuals with type 1 or type 2 diabetes, their caregivers, and health care providers managing diabetes) and familiar with the needs of the patients with diabetes. Project team members have conducted similar qualitative studies in the past and understand the health care access and resource disparity barriers (eg, education, transportation, and financial deficits) that exist for patients and providers living in underresourced, underrepresented rural and urban communities across Alaska, Idaho, and Virginia.

Approach

To ensure consistency in the data collection process, a moderator’s guide was developed to facilitate and standardize the focus groups and interviews. Guided by the established health technology assessment literature, the moderator’s guide scenarios and questions were developed and drafted by the research team and focused on (1) participant understanding of smart products and devices that use AI to manage diabetes, (2)

information needs to effectively and safely use AI/ML applications, and (3) participant suggestions on how best to communicate the necessary information to patients and providers to safely and effectively use applications and devices. For each application or device, we generated a patient-friendly description of the technology and how AI/ML was used. We generated context-specific queries for each example. Questions assessed patient and provider information needs, expected regulatory or other assurances, trust, and general perceptions of the value of the application. Scenarios were tested and refined during pilot sessions with a set of 4 patients and a provider. Questions were posed to providers in semistructured interviews that were similar to those asked of patients; the questions focused on information needed by patients to safely and effectively use AI/ML applications for diabetes management.

All focus group sessions and semistructured interviews were conducted by trained team personnel (RR, CL, and IW) who understand diabetes management challenges patients and providers face, know how to think about the problem (ie, reflexivity), and are sensitive as to how the data collection process may shape individual- and community-level responses (ie, research problem framing). This unique combination of professional experience, health training, and community engagement supported a more comprehensive understanding of training needs, sustainable training program development and implementation, and took into account prior assumptions, factors (ie, social contextual inquiry), and approaches used by the team (eg, diabetes and device information sharing) to overcome limited patient and provider AI/ML understanding and identify and recognize unmet information needs due to limited device and system experience [21].

RR, the qualitative research lead, conducted a 60-minute Zoom-based training session with all research team members to ensure focus group and interview consistency. Pilot training sessions were recorded, providing pertinent technology-based examples that focused on unmet patient and provider training needs (ie, use, maintenance, and troubleshooting), device safety concerns (alerts, warnings, and functionality), preferences for device testing, information sharing concerns, and other factors directly and indirectly related to device use (trust).

Ethical Approval

This study was granted expedited approval with a waiver of written consent (IRB-FY2021-259 for the work with patients and IRB-FY2021-260 for the work with providers) by the Idaho State University Institutional Review Board (IRB) and is subject to university research governance procedures. The Idaho State University IRB was also approved as the single IRB of record for the University of Virginia site. Participants or their legal guardians verbally consented to participation at the time of the interviews or focus group scheduling. Verbal consent was confirmed and documented again prior to the interviews or focus group initiation. All research was performed in accordance with relevant guidelines and regulations applicable to human subject participation and the Declaration of Helsinki.

Theoretical Framework

The Consolidated Framework for Implementation Research (CFIR) provides a menu of distinct constructs associated with effective program implementation (eg, implementation and organizational climate, culture, and context) and systematic analysis, and it supports incorporation of organization findings into practice [22,23]. Implementation climate, our primary construct, focuses on the impact that climate has on the implementation of innovative and progressive services, and the extent to which organization members perceive that an innovation is expected, supported, and rewarded by their organization or community [23-26].

Participant Selection

To recruit patients with type 1 or type 2 diabetes, flyers were distributed through local community groups, health care clinics, and diabetes educators. These groups included, but were not limited to, the Diabetes Alliance of Idaho, Camp Hodia, Idaho Primary Care Association, Community Council of Idaho, local community venues (churches and libraries), and local health care clinics (St. Luke's Endocrinology, Idaho Nutrition Associates, Idaho State University clinics, Full Circle Health, and University of Virginia [UVA] Health). The flyer was also shared with Facebook groups, including the Juvenile Diabetes Research Foundation Idaho, Native American Coalition of Boise, and Latter-Day Saints church groups. Lastly, the flyer was also promoted through paid promotion on Facebook. Paid promotion targeted the southern Idaho and Anchorage, Alaska, areas.

The flyer contained information regarding the study purpose, focus group eligibility, compensation, investigator contact information, and a screening survey link for interested individuals. The screening survey included full study details and collected eligibility and contact information. After individuals completed the screening survey, the project coordinator or research team member called them to confirm their interest in participation, reviewed consent, collected necessary information (ie, participant age, gender, diabetes diagnosis, race/ethnicity, technology use, and education level) and enrolled them. Participants could also complete the consent paperwork electronically or use paper forms, in person, before the focus group or interview. We used inclusive focus group methods to ensure participants' psychological safety and to encourage engagement. Two sessions had a majority of African American participants and 1 session had a majority of Native American/Alaska Native individuals.

Investigators used their relationship with area providers to recruit participants. In addition to these established relationships, area providers were also identified through an online search and contacted via email. We sought to recruit both physicians and certified diabetes care and education care specialists (CDCES) who care for patients with diabetes. After providers expressed their interest and willingness to participate in interviews, screening paperwork was completed, and their consent was verbally obtained prior to beginning the interview. We conducted semistructured interviews with providers using the established moderator's guide and Zoom, an online meeting platform. All focus group and interview sessions were

audio-recorded and transcribed. Individuals received a US \$75 gift card as an incentive for their participation.

Data Analysis

Grounded theory guided data gathering, synthesis, and analysis [27-29]. Thematic content and constant comparison analysis were used to identify relevant themes and allow for general and across-group assessments for both exploratory and verification purposes. An inductive approach was used to link data to key concepts, including patient perceptions of trust, value, accuracy, transparency, assurances, and preferred patient-provider approaches to application interaction [28,29]. QDA Miner (Provalis) [30] qualitative coding software was used for analysis. During the first stage of analysis, each transcript was systematically coded by at least two coders, with an initial codebook created based on moderator questions and initial review of the transcripts. Data were chunked into smaller units, definitions were established for each code, and the code/definition was attached to each unit (open coding). During the second stage, codes were grouped into categories (axial coding). Lastly, in the third stage, the researchers met frequently to refine and finalize codes (selective coding), identify discrepancies, achieve consensus, and establish the final codebook. Two coders systematically coded the data generating descriptive and analytic themes and identified patterns and dominant concepts that emerged during analysis. Where possible, codes associated with responders (ie, patient characteristics) were also included ([Multimedia Appendix 1](#)).

Representative quotes were sorted by codes, summary descriptions for each code were written, and information was linked to demographic data to identify additional patterns and themes. Preferred information or labeling presentation approaches and desired content were categorized and cross-referenced to patient classifications and themes were identified and prioritized. We used progressive analysis (data analysis concurrently with data collection) to support selection of scenarios and decisions on when enough sessions had been completed to achieve saturation in qualitative responses to key concepts [27,28,31]. Our full team of investigators reviewed (and iterated as needed) definitions, coding rules, and emerging themes (within the context of relevant interviewee quotes) for rigor, credibility, authenticity, sensitivity, and thoroughness [31]. The Consolidated Criteria for Reporting Qualitative Research (COREQ) were used to ensure comprehensive reporting of the qualitative data [32].

Results

General Characteristics

Between August and October 2022, we recruited and interviewed 41 patient participants ([Table 1](#)) to participate in 1 of 9 patient focus group sessions, 3 patient interview sessions (it should be noted that with teenagers, we conducted one-on-one sessions due to after-school conflicts), or 6 provider interviews. Provider interviews consisted of 3 pharmacists or CDCES, 2 primary care providers, and 1 diabetologist.

Table 1. Participant demographics (N=41).

Characteristics	Values
Age category, n (%)	
Adults (aged 20-89 years)	38 (93)
Teenagers (aged 16-19 years)	3 (7)
Age, (years), mean (SD)	48.4 (20.4)
Age, (years), median (IQR)	48 (32-66)
Gender^a, n (%)	
Male	19 (46)
Female	21 (51)
Diabetes type, n (%)	
Type 1	17 (41)
Type 2	24 (59)
Race^b, n (%)	
Alaska Native/American Indian	7 (17)
Black	13 (32)
White	24 (59)
Advanced technology user^a, n (%)	
Yes	33 (80)
No	7 (17)
Education level, n (%)	
Some high school	3 (7)
High school, General Educational Development test, or equivalent	6 (15)
Trade school, apprenticeship, or equivalent	4 (10)
Associate's degree	8 (20)
Bachelor's degree	9 (22)
Postgraduate or professional degree	11 (27)

^aData for 40 participants only; percentages are of 41 participants and do not add up to 100.

^bNot mutually exclusive groups; percentages do not add up to 100.

Themes, Subthemes, and Representative Quotes

Representative quotes are provided with relevant codes, themes, and subthemes: information needs ([Multimedia Appendix 2](#)), safety ([Multimedia Appendix 3](#)), and trust ([Multimedia Appendix 4](#)). Information needs were broken down into general needs, as well as training and informational support needs, preferences for information sharing, sources of information, troubleshooting, and information maintenance needs. Themes, subthemes, and representative quotes highlighted in [Multimedia Appendix 2](#) emphasized the importance of patient training and ready access to necessary information tools and resources, especially in response to AI/ML application alerts and warnings. Participants requested that information and training be provided in a number of different ways (eg, pamphlets, in-person training, computer-guided supports, and sharing of patient experiences). [Multimedia Appendix 3](#) presents safety concerns and needs identified by participants. Suggestions focused on input controls, alerts, reporting, override functions and manufacturer labeling,

information, and device mandates that could increase safety and improve AI/ML application trust. Lastly, [Multimedia Appendix 4](#) shows factors affecting participant trust and use of AI/ML applications. Reliability and accuracy of the measures in the specific population, AI/ML application limitations, and the impact of endorsements on trust are presented.

Discussion

Principal Findings

In health care, use of advanced computational methods and related AI/ML applications is expanding [1,2]. Provider- and patient-facing devices and applications (eg, continuous glucose monitors, insulin pumps, electronic health record-integrated decision supports, and mobile health apps) show great promise for improving diagnosis, data interpretation, and use of data to support treatment recommendations, dosage adjustment and management, and risk assessment [33].

While there is emerging research on public perceptions of responsible AI/ML application use, in general, little is known about how user interaction with specific AI/ML applications or related system information (eg, labels, intended use statements, and warnings) influences patient and provider perceptions of performance and addresses the ethical concerns or risks related to AI/ML use, especially in diabetes management and tailored medication therapy [2,6,34,35]. In order to provide useful guidance related to the representation of AI or AI-related explanations to patients with diabetes, it is important to explore patient and provider understanding of AI/ML applications, identify safety concerns with AI/ML use, and address underlying mistrust of AI/ML devices to support realistic contexts of use. In our research, we identified themes and subthemes and present summary descriptions, representative quotes, and relevant respondent data that identify and highlight the diverse patient and provider perspectives on unmet or suboptimal AI/ML application information and training needs, unaddressed safety concerns, and factors that influence patient and provider trust in the use of AI/ML applications for diabetes management.

Information and Training Needs

As we are all aware, diabetes is highly prevalent in the United States, affecting approximately 10% of Americans and 27% of people aged over 65 years [32]. The potential for AI/ML applications to improve outcomes for people living with diabetes is significant; however, information and training are necessary to support the human factors associated with safe and effective AI/ML application use in diabetes management, especially in older adults [35-37]. Patients need to understand all metrics displayed on the device to safely and effectively manage their diabetes. In our qualitative work, we found many patients rely on health care professionals as their primary resource for information about the appropriateness, quality, and safety of selected diabetes management technology. Most health care professionals may not have the necessary knowledge and experience with all available technology platforms to support meaningful use and troubleshooting of AI/ML applications for diabetes management; therefore, they require external support. In fact, according to a technology review conducted by the United Kingdom's National Health Service, rapid technological change requires that all health care providers (eg, doctors, nurses, pharmacists, and paramedics) receive extensive technology training [38].

This finding is consistent with the literature exploring patients' and health care professionals' perspectives toward technology use in diabetes management [39] and the concerns regarding safe and effective use of available technology that may be exacerbated if and when AI/ML applications become more available to patients (ie, over-the-counter and prescription applications) [40]. Therefore, it is essential that both patient and provider information and training needs are addressed to ensure patient diabetes management and safety needs are met by AI/ML device use (eg, understanding of device functionality, data availability, and safety functions). In fact, most participants in our study wanted and needed more information about the device or application than they initially received during training (eg, what it was measuring, why it was measuring it, and how results would be used to improve their health). Patients requested

that device information be clear, concise, and written in lay terms and that comprehensive information be provided in a number of different ways (eg, in-person training, hands-on device training, real-world instructional videos, manufacturer videos clips and targeted frequently asked questions, pamphlets, and cheat-sheets) to accommodate different learners and learning styles. Many patients also requested that peer-to-peer training and evidence-based informational resources be provided to support real-life device use and troubleshooting. We also found that the amount of information provided at any one time was often a limiting factor and was both overwhelming and confusing to the patients and caregivers. It is important to note that initially, patients in our study were unsure of their own information needs, and that questions arose with daily device and application use over the following weeks. This suggests that a tiered or layered approach to teaching [41], validated and used in adult learning and education models, be included. Maintenance, troubleshooting, and potentially life-threatening alerts might be necessary to ensure appropriate and safe device use. A number of patients and providers in our study suggested a tiered approach to both knowledge assessment and functionality, which would require a minimal level of disease state and device or application knowledge to allow users to enable specific functions. The staged or tiered approach to training was viewed by many patients as an effective and efficient training mechanism aligned with patient understanding. The ability to watch instructions in segments was thought to allow for device mastery. Patients also requested the ability to trial a number of devices and to be connected to all relevant systems to ensure that the device is appropriate for them (eg, considering type of diabetes and experience with technology). This is consistent with patient training needs and requests seen in literature regarding human factors and usability engineering for medical device labeling and function, especially among older adults [36,39,42].

Lastly, there were a number of participant suggestions regarding training and support that could be provided by device manufacturers to improve device use and testing. Suggestions included the following: (1) provide a basic starter guide for the first few days of use; (2) provide practice devices that allow for hands-on trials; (3) provide links to online resources, local supports, and reputable community resources (eg, professional organizations, blogs, and personal reviews) on the manufacturer website; (4) provide 24/7 live in-person or virtual emergency support; and (5) provide brief, searchable, instructional resources, such as videos indexed by problem and answers to frequently asked questions.

Safety

In respect to safety, patients in our study were most concerned with (1) having a clear understanding of alerts and warnings, (2) being able to recognize and rapidly respond to a potentially life-threatening situation (eg, device overrides, function lockdowns, and system-down alerts), (3) knowing immediately if there were device connectivity issues that impede overall diabetes management (eg, the continuous glucose monitor not connected to the insulin pump), and (4) having safeguards to reduce the risk of user error (eg, data field restrictions and order entry confirmation requirements).

Participants wanted access to real-time, live device safety support offering them the ability to more effectively and efficiently troubleshoot issues with devices that directly control insulin delivery. Participants also voiced concerns regarding the number of alerts they received, the alert descriptions being provided as codes, the information provided by the manufacturer or provider about what to do to address the alert (device instructions), and mechanisms in place to stop alerts once the patients has addressed them (to avoid alert fatigue). This is consistent with the scientific and lay literature; having clear predictive and real-time alerts is important but so is ensuring that alerts can be tailored to patient needs and address provider concerns [43-45].

Providers stressed the importance of patients having access to a limited number of clear, clinically important alerts and necessary alarms and the provision of patient education focused on understanding what to do in the case of an alert or alarm. If users cannot see or interpret the alert, they will not respond appropriately, a documented challenge for many older adults [37,46]. In order for required safety information provided to patients to be useful, it needs to be immediate, detailed, and prescriptive and provide simple instructions to the patient and caregiver [47,48]. It is also important that device updates related to safety and device functionality be pushed out automatically to ensure continued safe and effective device and application use. Lastly, it was recommended by participants that all safety features need to either remind or directly connect patients to providers, emergency services (eg, 911 and Medic Alert), and necessary troubleshooting resources to help support patient understanding and encourage patient ownership of care.

Trust

Trust in the device or application was based on trust in the health care provider's recommendations and the participant's experience with that health care provider; however, it also extended beyond the clinical interface to the collection, collation, and use of personal data [49-53]. In our study, individuals consistently treated by the same health care provider or specialist appeared to have more trust in the provider-recommended device. However, it is important to note that concerns regarding blind trust were voiced by a number of patients and providers in our study and that trust in the device was directly related to patient experience, device accuracy, and duration of device use.

AI/ML application use can be associated with a number of risks as well as benefits. As such, our findings are supported by other research that emphasizes the complexity of and need for trust being embedded in all aspects of AI. Specifically, Lockey et al [50] support this finding, showing that transparency, explainability, and accuracy metrics are important, although

they may not be sufficient, to garner trust in AI applications. In line with our methodological approach, Lockey and colleagues [50] also suggest the need to examine multiple key stakeholders in relation to AI systems and their varying expectations and alignment with the outcomes of using the AI device.

Participants expressed the need for exposure to the device and a mechanism in place to double-check readings and functionality to build trust; they also expressed to need for the opportunity to question device results and troubleshoot concerns with providers and other health care team members. Participants raised an important point on having detailed and accessible information on the population characteristics (ie, age, race/ethnicity, gender, and diabetes type) of those who tested the device or application. Participants wanted to know that the device was tested in individuals similar to them. These results are in line with best practices for ensuring and promoting trust in AI implementation, such as including representative and equitable populations in its development, having a user-centered design, and ensuring constant accountability of the algorithm being used to maintain accuracy [51]. Given the importance of human factors and the associated patient outcomes in use of AI devices, it is essential to understand how trust is linked to the needs of the user and design requirements [52,53]. Our data support optimizing the opinions of patients and users and acknowledging that trust shapes clinicians' and patients' use and initial adoption of AI devices [52].

The implementation of the strategies discussed above can increase proper use, safety, and trust regarding AI-enabled medical devices. In an informal review of patient-facing AI systems available from the FDA [54], we found that current apps and systems lack detailed information and resources for users, both patients and providers. We believe this makes our findings even more important. As manufacturers and device makers hopefully integrate our suggestions, real-world examples will arise. Further investigation will then be needed to optimize AI system interfaces.

Conclusions and Next Steps

Our work supplements the emerging literature related to public perceptions of responsibility and ethics in AI/ML device and application use [7,13,14]. We hope that our findings inform the FDA's decisions on public health and safety related to AI/ML devices and applications. AI/ML applications demonstrate a great deal of promise; however, even greater outcomes will be realized if ethical and responsible AI design engenders greater engagement and use by all. It is important to understand how to present information to patients about AI/ML characteristics identified as important to them, such as data privacy, fairness, accuracy, and risks.

Acknowledgments

We would like to thank the many individuals and organizations that assisted in this research, including Karalynn Jensen, John Holmes, Viola Holmes, the groups that helped in recruitment of participants, and those at the Food and Drug Administration (FDA). We are grateful to the participants who spent their valuable time with us to share their thoughts and experiences.

This work was supported by the Food and Drug Administration (FDABAA-21-00123). The content is solely the responsibility of the authors and does not necessarily represent the official views of the FDA.

Conflicts of Interest

None declared.

Multimedia Appendix 1

Codebook.

[[DOCX File , 18 KB - ai_v2i1e46487_app1.docx](#)]

Multimedia Appendix 2

Themes, subthemes, and representative quotes related to information needs.

[[DOCX File , 33 KB - ai_v2i1e46487_app2.docx](#)]

Multimedia Appendix 3

Themes, subthemes, and representative quotes related to safety.

[[DOCX File , 21 KB - ai_v2i1e46487_app3.docx](#)]

Multimedia Appendix 4

Themes, subthemes, and representative quotes related to trust.

[[DOCX File , 20 KB - ai_v2i1e46487_app4.docx](#)]

References

1. Yu K, Beam AL, Kohane IS. Artificial intelligence in healthcare. *Nat Biomed Eng* 2018 Oct 10;2(10):719-731. [doi: [10.1038/s41551-018-0305-z](#)] [Medline: [31015651](#)]
2. Contreras I, Vehi J. Artificial intelligence for diabetes management and decision support: Literature review. *J Med Internet Res* 2018 May 30;20(5):e10775 [FREE Full text] [doi: [10.2196/10775](#)] [Medline: [29848472](#)]
3. Rejala G, Ravi A, Churiwala S. Machine learning definition and basics. In: *An Introduction to Machine Learning*. Cham, Switzerland: Springer; 2019.
4. Wang H, Ma C, Zhou L. A brief review of machine learning and its application. 2009 Presented at: International Conference on Information Engineering and Computer Science; December 19-20, 2009; Wuhan, China. [doi: [10.1109/iciecs.2009.5362936](#)]
5. Salgado T, Tavares J, Oliveira T. Drivers of mobile health acceptance and use from the patient perspective: Survey study and quantitative model development. *JMIR Mhealth Uhealth* 2020 Jul 09;8(7):e17588 [FREE Full text] [doi: [10.2196/17588](#)] [Medline: [32673249](#)]
6. Schimmer R, Orre C, Öberg U, Danielsson K, Hörnsten Å. Digital person-centered self-management support for people with type 2 diabetes: Qualitative study exploring design challenges. *JMIR Diabetes* 2019 Sep 19;4(3):e10702 [FREE Full text] [doi: [10.2196/10702](#)] [Medline: [31538941](#)]
7. Rigby M. Ethical dimensions of using artificial intelligence in health care. *AMA J Ethics* 2019;21(2):E121-E124 [FREE Full text] [doi: [10.1001/AMAJETHICS.2019.121](#)]
8. Kumar P, Dwivedi YK, Anand A. Responsible artificial intelligence (AI) for value formation and market performance in healthcare: The mediating role of patient's cognitive engagement. *Inf Syst Front* 2021 Apr 29;1-24 [FREE Full text] [doi: [10.1007/s10796-021-10136-6](#)] [Medline: [33948105](#)]
9. 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](#)]
10. Romiti S, Vinciguerra M, Saade W, Anso Cortajarena I, Greco E. Artificial intelligence (AI) and cardiovascular diseases: An unexpected alliance. *Cardiol Res Pract* 2020;2020:4972346 [FREE Full text] [doi: [10.1155/2020/4972346](#)] [Medline: [32676206](#)]
11. Graham S, Depp C, Lee EE, Nebeker C, Tu X, Kim H, et al. Artificial intelligence for mental health and mental illnesses: An overview. *Curr Psychiatry Rep* 2019 Nov 07;21(11):116 [FREE Full text] [doi: [10.1007/s11920-019-1094-0](#)] [Medline: [31701320](#)]
12. Habli I, Lawton T, Porter Z. Artificial intelligence in health care: accountability and safety. *Bull World Health Organ* 2020 Apr 01;98(4):251-256 [FREE Full text] [doi: [10.2471/BLT.19.237487](#)] [Medline: [32284648](#)]
13. Ho A. Are we ready for artificial intelligence health monitoring in elder care? *BMC Geriatr* 2020 Sep 21;20(1):358 [FREE Full text] [doi: [10.1186/s12877-020-01764-9](#)] [Medline: [32957946](#)]
14. Digital Health Center of Excellence. Food and Drug Administration. URL: <https://www.fda.gov/medical-devices/digital-health-center-excellence/> [accessed 2023-05-19]

15. Artificial Intelligence and Machine Learning in Software as a Medical Device. Food and Drug Administration. URL: <https://tinyurl.com/rwrh739a> [accessed 2023-05-19]
16. Some digital divides persist between rural, urban and suburban America. Pew Research Center. URL: <https://tinyurl.com/2xyk728x> [accessed 2023-05-19]
17. Barriers to Telehealth in Rural Areas. Rural Health Information Hub. URL: <https://www.ruralhealthinfo.org/toolkits/telehealth/1/barriers> [accessed 2023-05-19]
18. Singh RP, Hom GL, Abramoff MD, Campbell JP, Chiang MF, AAO Task Force on Artificial Intelligence. Current challenges and barriers to real-world artificial intelligence adoption for the healthcare system, provider, and the patient. *Transl Vis Sci Technol* 2020 Aug;9(2):45 [FREE Full text] [doi: [10.1167/tvst.9.2.45](https://doi.org/10.1167/tvst.9.2.45)] [Medline: [32879755](https://pubmed.ncbi.nlm.nih.gov/32879755/)]
19. Kelly CJ, Karthikesalingam A, Suleyman M, Corrado G, King D. Key challenges for delivering clinical impact with artificial intelligence. *BMC Med* 2019 Oct 29;17(1):195. [doi: [10.1186/s12916-019-1426-2](https://doi.org/10.1186/s12916-019-1426-2)] [Medline: [31665002](https://pubmed.ncbi.nlm.nih.gov/31665002/)]
20. Barriers to AI in Healthcare. GreenBook. URL: <https://tinyurl.com/mr288s94> [accessed 2023-05-19]
21. Gemert-Pijnen LV, Kelders S, Kip H, Sanderman R, editors. *eHealth Research, Theory and Development*. London, UK: Routledge; 2018.
22. Damschroder L, Hall C, Gillon L, Reardon C, Kelley C, Sparks J, et al. The Consolidated Framework for Implementation Research (CFIR): progress to date, tools and resources, and plans for the future. *Implementation Sci* 2015 Aug 14;10(S1):A12 [FREE Full text] [doi: [10.1186/1748-5908-10-s1-a12](https://doi.org/10.1186/1748-5908-10-s1-a12)]
23. Weiner BJ, Belden CM, Bergmire DM, Johnston M. The meaning and measurement of implementation climate. *Implement Sci* 2011 Jul 22;6(1):78 [FREE Full text] [doi: [10.1186/1748-5908-6-78](https://doi.org/10.1186/1748-5908-6-78)] [Medline: [21781328](https://pubmed.ncbi.nlm.nih.gov/21781328/)]
24. Damschroder LJ, Aron DC, Keith RE, Kirsh SR, Alexander JA, Lowery JC. Fostering implementation of health services research findings into practice: a consolidated framework for advancing implementation science. *Implement Sci* 2009 Aug 07;4:50 [FREE Full text] [doi: [10.1186/1748-5908-4-50](https://doi.org/10.1186/1748-5908-4-50)] [Medline: [19664226](https://pubmed.ncbi.nlm.nih.gov/19664226/)]
25. Livet M, Haines ST, Curran GM, Seaton TL, Ward CS, Sorensen TD, et al. Implementation science to advance care delivery: A primer for pharmacists and other health professionals. *Pharmacotherapy* 2018 May;38(5):490-502. [doi: [10.1002/phar.2114](https://doi.org/10.1002/phar.2114)] [Medline: [29624704](https://pubmed.ncbi.nlm.nih.gov/29624704/)]
26. Powell BJ, Waltz TJ, Chinman MJ, Damschroder LJ, Smith JL, Matthieu MM, et al. A refined compilation of implementation strategies: results from the Expert Recommendations for Implementing Change (ERIC) project. *Implement Sci* 2015 Feb 12;10(1):21 [FREE Full text] [doi: [10.1186/s13012-015-0209-1](https://doi.org/10.1186/s13012-015-0209-1)] [Medline: [25889199](https://pubmed.ncbi.nlm.nih.gov/25889199/)]
27. Birks M, Mills J. *Grounded Theory: A Practical Guide*. Second edition. Thousand Oaks, CA: SAGE; 2015.
28. Charmaz K. *Constructing Grounded Theory*. 2nd Edition. Thousand Oaks, CA: SAGE; 2014.
29. Krueger RA, Casey MA. *Focus Groups: A Practical Guide for Applied Research*. 5th Edition. Thousand Oaks, CA: SAGE; 2015.
30. Miles MB, Huberman AM, Saldaña J. *Qualitative Data Analysis: A Methods Sourcebook*. Fourth Edition. Thousand Oaks, CA: SAGE; 2020.
31. Whittemore R, Chase SK, Mandle CL. Validity in qualitative research. *Qual Health Res* 2001 Jul;11(4):522-537. [doi: [10.1177/104973201129119299](https://doi.org/10.1177/104973201129119299)] [Medline: [11521609](https://pubmed.ncbi.nlm.nih.gov/11521609/)]
32. Tong A, Sainsbury P, Craig J. Consolidated criteria for reporting qualitative research (COREQ): a 32-item checklist for interviews and focus groups. *Int J Qual Health Care* 2007 Dec;19(6):349-357 [FREE Full text] [doi: [10.1093/intqhc/mzm042](https://doi.org/10.1093/intqhc/mzm042)] [Medline: [17872937](https://pubmed.ncbi.nlm.nih.gov/17872937/)]
33. Greenwood DA, Gee PM, Fatkin KJ, Peeples M. A systematic review of reviews evaluating technology-enabled diabetes self-management education and support. *J Diabetes Sci Technol* 2017 Sep;11(5):1015-1027 [FREE Full text] [doi: [10.1177/1932296817713506](https://doi.org/10.1177/1932296817713506)] [Medline: [28560898](https://pubmed.ncbi.nlm.nih.gov/28560898/)]
34. Nomura A, Noguchi M, Kometani M, Furukawa K, Yoneda T. Artificial intelligence in current diabetes management and prediction. *Curr Diab Rep* 2021 Dec 13;21(12):61 [FREE Full text] [doi: [10.1007/s11892-021-01423-2](https://doi.org/10.1007/s11892-021-01423-2)] [Medline: [34902070](https://pubmed.ncbi.nlm.nih.gov/34902070/)]
35. Singla R, Singla A, Gupta Y, Kalra S. Artificial intelligence/machine learning in diabetes care. *Indian J Endocrinol Metab* 2019;23(4):495-497 [FREE Full text] [doi: [10.4103/ijem.IJEM_228_19](https://doi.org/10.4103/ijem.IJEM_228_19)] [Medline: [31741913](https://pubmed.ncbi.nlm.nih.gov/31741913/)]
36. Liberman A, Buckingham B, Phillip M. Diabetes technology and the human factor. *Int J Clin Pract Suppl* 2011 Feb;65(170):83-90 [FREE Full text] [doi: [10.1111/j.1742-1241.2010.02583.x](https://doi.org/10.1111/j.1742-1241.2010.02583.x)] [Medline: [21323817](https://pubmed.ncbi.nlm.nih.gov/21323817/)]
37. Toschi E, Munshi MN. Benefits and challenges of diabetes technology use in older adults. *Endocrinol Metab Clin North Am* 2020 Mar;49(1):57-67 [FREE Full text] [doi: [10.1016/j.ecl.2019.10.001](https://doi.org/10.1016/j.ecl.2019.10.001)] [Medline: [31980121](https://pubmed.ncbi.nlm.nih.gov/31980121/)]
38. Vogel L. Doctors need retraining to keep up with technological change. *CMAJ* 2018 Jul 30;190(30):E920 [FREE Full text] [doi: [10.1503/cmaj.109-5637](https://doi.org/10.1503/cmaj.109-5637)] [Medline: [30061332](https://pubmed.ncbi.nlm.nih.gov/30061332/)]
39. Jain SR, Sui Y, Ng CH, Chen ZX, Goh LH, Shorey S. Patients' and healthcare professionals' perspectives towards technology-assisted diabetes self-management education. A qualitative systematic review. *PLoS One* 2020;15(8):e0237647 [FREE Full text] [doi: [10.1371/journal.pone.0237647](https://doi.org/10.1371/journal.pone.0237647)] [Medline: [32804989](https://pubmed.ncbi.nlm.nih.gov/32804989/)]
40. The 7 Diabetes Devices That May Be Available Soon. GoodRx. URL: <https://www.goodrx.com/conditions/diabetes/diabetes-devices> [accessed 2023-05-19]
41. Adult Learning and Education System Building Approach (ALESBA). DVV International. URL: <https://tinyurl.com/4trcycfx> [accessed 2023-05-19]

42. Applying Human Factors and Usability Engineering to Medical Devices. Food and Drug Administration. URL: <https://tinyurl.com/24ae6are> [accessed 2023-05-19]
43. How to reduce diabetes alerts while making the most of our technology. Integrated Diabetes Services. URL: <https://integrateddiabetes.com/how-to-reduce-diabetes-alerts-while-making-the-most-of-our-technology/> [accessed 2023-05-19]
44. Abraham SB, Arunachalam S, Zhong A, Agrawal P, Cohen O, McMahon CM. Improved real-world glycemic control with continuous glucose monitoring system predictive alerts. *J Diabetes Sci Technol* 2021 Jan;15(1):91-97 [FREE Full text] [doi: [10.1177/1932296819859334](https://doi.org/10.1177/1932296819859334)] [Medline: [31272204](https://pubmed.ncbi.nlm.nih.gov/31272204/)]
45. Dave D, Erraguntla M, Lawley M, DeSalvo D, Haridas B, McKay S, et al. Improved low-glucose predictive alerts based on sustained hypoglycemia: Model development and validation study. *JMIR Diabetes* 2021 Apr 29;6(2):e26909 [FREE Full text] [doi: [10.2196/26909](https://doi.org/10.2196/26909)] [Medline: [33913816](https://pubmed.ncbi.nlm.nih.gov/33913816/)]
46. Keller SC, Gurses AP, Werner N, Hohl D, Hughes A, Leff B, et al. Older adults and management of medical devices in the home: Five requirements for appropriate use. *Popul Health Manag* 2017 Aug;20(4):278-286 [FREE Full text] [doi: [10.1089/pop.2016.0070](https://doi.org/10.1089/pop.2016.0070)] [Medline: [28075698](https://pubmed.ncbi.nlm.nih.gov/28075698/)]
47. Lee L, Maher ML. Factors affecting the initial engagement of older adults in the use of interactive technology. *Int J Environ Res Public Health* 2021 Mar 11;18(6):2847 [FREE Full text] [doi: [10.3390/ijerph18062847](https://doi.org/10.3390/ijerph18062847)] [Medline: [33799568](https://pubmed.ncbi.nlm.nih.gov/33799568/)]
48. Tsai T, Lin W, Chang Y, Chang P, Lee M. Technology anxiety and resistance to change behavioral study of a wearable cardiac warming system using an extended TAM for older adults. *PLoS One* 2020;15(1):e0227270 [FREE Full text] [doi: [10.1371/journal.pone.0227270](https://doi.org/10.1371/journal.pone.0227270)] [Medline: [31929560](https://pubmed.ncbi.nlm.nih.gov/31929560/)]
49. Richards B, Scheibner J. Health technology and big data: Social licence, trust and the law. *J Law Med* 2022 Jun;29(2):388-399. [Medline: [35819379](https://pubmed.ncbi.nlm.nih.gov/35819379/)]
50. Lockey S, Gillespie N, Holm D, Someh I. A review of trust in artificial intelligence: Challenges, vulnerabilities and future directions. 2021 Presented at: 54th Hawaii International Conference on System Sciences; January 5-8, 2021; Maui, HI. [doi: [10.24251/hicss.2021.664](https://doi.org/10.24251/hicss.2021.664)]
51. Roski J, Maier EJ, Vigilante K, Kane EA, Matheny ME. Enhancing trust in AI through industry self-governance. *J Am Med Inform Assoc* 2021 Jul 14;28(7):1582-1590 [FREE Full text] [doi: [10.1093/jamia/ocab065](https://doi.org/10.1093/jamia/ocab065)] [Medline: [33895824](https://pubmed.ncbi.nlm.nih.gov/33895824/)]
52. Asan O, Bayrak AE, Choudhury A. Artificial intelligence and human trust in healthcare: Focus on clinicians. *J Med Internet Res* 2020 Jun 19;22(6):e15154 [FREE Full text] [doi: [10.2196/15154](https://doi.org/10.2196/15154)] [Medline: [32558657](https://pubmed.ncbi.nlm.nih.gov/32558657/)]
53. Asan O, Choudhury A. Research trends in artificial intelligence applications in human factors health care: Mapping review. *JMIR Hum Factors* 2021 Jun 18;8(2):e28236 [FREE Full text] [doi: [10.2196/28236](https://doi.org/10.2196/28236)] [Medline: [34142968](https://pubmed.ncbi.nlm.nih.gov/34142968/)]
54. Artificial Intelligence and Machine Learning (AI/ML)-Enabled Medical Devices. Food and Drug Administration. URL: <https://tinyurl.com/eyzd4rdb> [accessed 2023-05-19]

Abbreviations

AI: artificial intelligence
CDCES: certified diabetes care and education specialist
CFIR: Consolidated Framework for Implementation Research
COREQ: Consolidated Criteria for Reporting Qualitative Research
FDA: Food and Drug Administration
IRB: institutional review board
ML: machine learning
PEAC: Patient Engagement Advisory Committee
UVA: University of Virginia

Edited by K El Emam, B Malin; submitted 13.02.23; peer-reviewed by N Jiwani, V Ochs; comments to author 22.04.23; revised version received 10.05.23; accepted 14.05.23; published 19.06.23.

Please cite as:

Robinson R, Liday C, Lee S, Williams IC, Wright M, An S, Nguyen E
Artificial Intelligence in Health Care—Understanding Patient Information Needs and Designing Comprehensible Transparency: Qualitative Study
JMIR AI 2023;2:e46487
URL: <https://ai.jmir.org/2023/1/e46487>
doi: [10.2196/46487](https://doi.org/10.2196/46487)
PMID: [38333424](https://pubmed.ncbi.nlm.nih.gov/38333424/)

©Renee Robinson, Cara Liday, Sarah Lee, Ishan C Williams, Melanie Wright, Sungjoon An, Elaine Nguyen. Originally published in JMIR AI (<https://ai.jmir.org>), 19.06.2023. This is an open-access article distributed under the terms of the Creative Commons Attribution License (<https://creativecommons.org/licenses/by/4.0/>), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR AI, is properly cited. The complete bibliographic information, a link to the original publication on <https://www.ai.jmir.org/>, as well as this copyright and license information must be included.

Original Paper

Artificial Intelligence–Enabled Software Prototype to Inform Opioid Pharmacovigilance From Electronic Health Records: Development and Usability Study

Alfred Sorbello^{1*}, DO, MPH; Syed Arefinul Haque^{1*}, PhD; Rashedul Hasan^{1*}, PhD; Richard Jermyn^{2*}, DO; Ahmad Hussein^{2*}, MBS; Alex Vega^{2*}, BS; Krzysztof Zembruski^{2*}, BA; Anna Ripple^{3*}, MLS; Mitra Ahadpour^{1*}, MD

¹Center for Drug Evaluation and Research, US Food and Drug Administration, Silver Spring, MD, United States

²Neuromuscular Institute, Rowan-Virtua School of Osteopathic Medicine, Stratford, NJ, United States

³Lister Hill National Center for Biomedical Communications, National Library of Medicine–National Institutes of Health, Rockville, MD, United States

*all authors contributed equally

Corresponding Author:

Rashedul Hasan, PhD

Center for Drug Evaluation and Research

US Food and Drug Administration

10903 New Hampshire Avenue

Silver Spring, MD, 20993

United States

Phone: 1 (888) 463 6332

Email: mdrashedul.hasan@fda.hhs.gov

Abstract

Background: The use of patient health and treatment information captured in structured and unstructured formats in computerized electronic health record (EHR) repositories could potentially augment the detection of safety signals for drug products regulated by the US Food and Drug Administration (FDA). Natural language processing and other artificial intelligence (AI) techniques provide novel methodologies that could be leveraged to extract clinically useful information from EHR resources.

Objective: Our aim is to develop a novel AI-enabled software prototype to identify adverse drug event (ADE) safety signals from free-text discharge summaries in EHRs to enhance opioid drug safety and research activities at the FDA.

Methods: We developed a prototype for web-based software that leverages keyword and trigger-phrase searching with rule-based algorithms and deep learning to extract candidate ADEs for specific opioid drugs from discharge summaries in the Medical Information Mart for Intensive Care III (MIMIC III) database. The prototype uses MedSpacy components to identify relevant sections of discharge summaries and a pretrained natural language processing (NLP) model, Spark NLP for Healthcare, for named entity recognition. Fifteen FDA staff members provided feedback on the prototype's features and functionalities.

Results: Using the prototype, we were able to identify known, labeled, opioid-related adverse drug reactions from text in EHRs. The AI-enabled model achieved accuracy, recall, precision, and F_1 -scores of 0.66, 0.69, 0.64, and 0.67, respectively. FDA participants assessed the prototype as highly desirable in user satisfaction, visualizations, and in the potential to support drug safety signal detection for opioid drugs from EHR data while saving time and manual effort. Actionable design recommendations included (1) enlarging the tabs and visualizations; (2) enabling more flexibility and customizations to fit end users' individual needs; (3) providing additional instructional resources; (4) adding multiple graph export functionality; and (5) adding project summaries.

Conclusions: The novel prototype uses innovative AI-based techniques to automate searching for, extracting, and analyzing clinically useful information captured in unstructured text in EHRs. It increases efficiency in harnessing real-world data for opioid drug safety and increases the usability of the data to support regulatory review while decreasing the manual research burden.

(JMIR AI 2023;2:e45000) doi:[10.2196/45000](https://doi.org/10.2196/45000)

KEYWORDS

electronic health records; pharmacovigilance; artificial intelligence; real world data; EHR; natural language; software application; drug; Food and Drug Administration; deep learning

Introduction

Postmarketing drug safety surveillance at the Center for Drug Evaluation and Research (CDER) of the US Food and Drug Administration (FDA) aims to detect, characterize, monitor, and prevent adverse drug reactions (ADRs) for FDA-approved drugs and therapeutic biologic products. Biomedical resources used to detect adverse drug event (ADE) safety signals include clinical trials, spontaneous adverse event (AE) reports submitted to the FDA Adverse Events Reporting System (FAERS), published scientific reports in the literature, and others. The FAERS database compiles AE and medication error reports submitted to the FDA to support postmarket drug safety monitoring. FAERS monitoring has yielded information on rare ADEs, but the information is limited by underreporting [1,2]. Multimodal approaches to pharmacovigilance using multiple biomedical resources may offer improved drug safety signal detection compared to reliance on single resources [3].

Electronic health records (EHRs) are a rich source of real-world information that may potentially serve as a new complementary drug safety resource. Although not specifically created to document ADEs, the EHR may provide information about product side effects, including those that occur a prolonged time following initial drug exposure [4], and may contribute to assessments of the safety of generic and pediatric drug products [5,6]. EHRs have been explored to complement ADE signal identification from spontaneous AE reports [7].

Published scientific reports describe various natural language processing (NLP) and artificial intelligence (AI)-based approaches to analyzing text from EHRs for ADE detection and pharmacovigilance. Named entity recognition (NER) to identify drug and AE mentions in text followed by extraction of the relationships between those entities is a critical technical challenge in building successful analytical algorithms. In general, keywords, rule-based algorithms, and machine learning methods have been used for case detection [8]. Some early studies used trigger phrases to screen the text of discharge summaries for AE concepts [9,10]. Established NLP algorithms applied to AE detection include MedLEE, which identifies clinical concepts and cross-maps them to Unified Medical Language System (UMLS) concepts [11]; MetaMap, which processes biomedical text and maps it to the UMLS [12]; and Clinical Text Analysis and Knowledge Extraction System (cTAKES), an NLP system that incorporates rules and machine learning [13]. More recent studies use multiple NLP models, including long short-term memory (LSTM), conditional random field (CRF), support vector machines (SVMs), and bidirectional encoder representations from transformers (BERTs) [14]. Shared task challenges designed to promote advances in NLP for drug safety and ADE detection from EHRs have been conducted in recent years, including the MADE 1.0 challenge [15] and the n2c2 Clinical Challenge [16]. Text analytic engines, such as Amazon Comprehend Medical, Microsoft Text Analytics for Health, and the Google Healthcare Natural Language application programming interface, are deep learning-based pretrained models. These models can perform a variety of general health care NLP tasks, such as NER, relation detection, entity disambiguation, and others [17]. We combine a similar deep

learning model with domain-specific, rule-based algorithms from domain expertise to detect opioid-related ADEs (ORADEs) from clinical notes.

Using novel AI methods, time-consuming manual chart review can be automated to provide active surveillance with enhanced detection of emerging product safety issues in near-real time. Opioids are one of the most frequently implicated drug classes for ADRs in hospitalized patients and are associated with confusion, constipation, respiratory depression, sedation, ileus, hypotension, and other ADRs [18]. One study reported an ORADE prevalence rate of 9.1% in previously opioid-free surgical patients [19]. In this manuscript, we report on the development of and user feedback for SPINEL (Supporting Pharmacovigilance by Leveraging Artificial Intelligence Methods to Analyze Electronic Health Records Data), a novel AI-enabled software prototype that analyzes unstructured text in discharge summaries to extract candidate ADEs for opioid drugs. FDA participants provide feedback on the serviceability of the prototype in meeting their needs to support drug safety, research, and regulatory decision-making.

Methods

Ethical Considerations

This study does not meet the requirements of research involving human subjects as defined by the US Department of Health and Human Services (45CFR46) for the following reasons: (1) there was no interaction or intervention with human subjects; (2) MIMIC is a free, publicly available database and the authors have completed the required Collaborative Institutional Training Initiative training and data use agreement; (3) all MIMIC III data were deidentified in accordance with Health Insurance Portability and Accountability Act requirements, including removal of 18 identifying data elements; (4) protected health information has been removed from free text fields; and (5) no personally identifiable information was available to the study investigators.

Data Source

EHR Data

We limited our work to publicly accessible EHR databases and focused on the free text in discharge summaries from the Medical Information Mart for Intensive Care III (MIMIC III) [20]. This database contains EHRs from 2001 through 2012 from a single health care center; the records are encoded with codes in the International Classification of Diseases, Ninth Revision (ICD-9). We leveraged ICD-9 code E935.2, which indicates opioids and other narcotics causing AEs in therapeutic use, to prescreen discharge summaries that may contain information on ORADEs. We identified 227 summaries consisting of 227 unique hospital-event records for 226 unique patients. We planned to explore the more recently released MIMIC IV EHR database for additional cases, but the discharge summaries were not made publicly accessible until after this project was completed.

Reference Data Set for Testing and Training

Considering that ICD-9 codes have limited positive predictive value for drug safety surveillance [21], 2 medical students (AV, KZ) and a physician (AS) conducted independent manual reviews of the 227 discharge summaries identified by ICD-9 prescreening (as above) to manually assess for documentation of ORADEs in the text. We did not use a formal annotation guideline; positive assessments were based on specific textual mentions describing opioid drug exposure and adverse events either linked or potentially linked to the exposure irrespective of the severity or seriousness of the events. To create a reference data set of discharge summaries with true positive and negative cases, positive assessment for an ORADE required agreement among all 3 reviewers. Discrepancies were reconciled through joint discussion. The 3 reviewers had similar assessments for

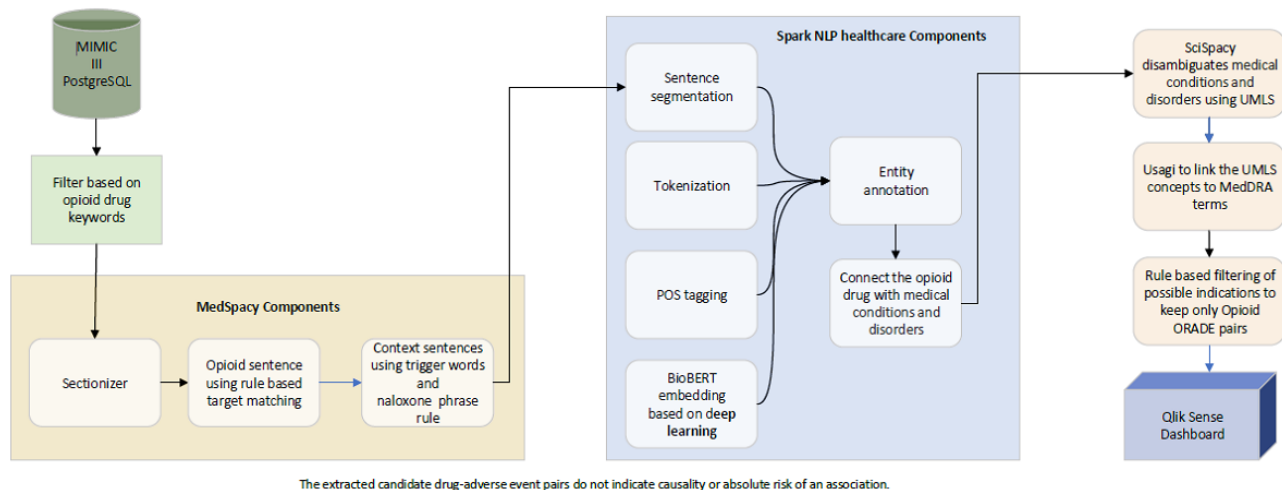
ORADE documentation for 174 (77%) of the 227 discharge summaries reviewed. We trained our AI-enabled model on 181 (80%) of the discharge summaries and used the remaining 46 (20%) for testing.

NLP Process

Detection of Sections in Discharge Summaries

Based on a manual review, we identified 3 sections with the highest frequency of ORADE mentions: “brief hospital course,” “hospital course,” and “history of present illness.” In our AI-enabled model (Figure 1), we used the Sectionizer module in the MedSpacy open-source Python library [22] to automate the identification of those component sections in the sample of discharge summaries.

Figure 1. The artificial intelligence-enabled model is depicted with natural language processing and rule-based algorithms, MedSpacy sectionizer components, Spark NLP for Healthcare entity recognition components, SciSpacy disambiguation of terms, Usagi interconnection of UMLS concepts with MedDRA terminology, and further filtering of ORADE pairs. A higher resolution version of this figure is available in [Multimedia Appendix 1](#). MIMIC: Medical Information Mart for Intensive Care; NLP: natural language processing; POS: part of speech; UMLS: Unified Medical Language System; MedDRA: Medical Dictionary for Regulatory Activities; ORADE: opioid-related adverse drug event.



Identifying ORADE Context Sentences Using Keywords, Trigger Phrases, and Rule-Based Algorithms

Using MedSpacy components, we divided the unstructured text in the 3 component sections of the discharge summaries into individual sentences. We identified the context sentences in 2 stages. In the first stage, we identified the sentences that contained one or more mentions of opioid-drug generic terms or opioid-drug brand names using keyword lists manually constructed by one of the team members (AS). The drug names were aligned with RxNorm terminology.

In the second stage, we used 2 rule-based approaches to identify context sentences with mentions of possible ORADEs. First, the trigger-phrase rule: We applied trigger phrases [23] to link mentions of an opioid drug with ADE terms using the MedSpacy context algorithm [24]. We curated 58 additional trigger phrases (Multimedia Appendix 2) from the training subset of the reference data set and included them in our analysis. To capture mentions of opioid drugs and ADEs that did not co-occur in the same sentence, we searched for those terms in the 3 sentences preceding and following the sentence of interest based on reported heuristics [23].

An example of a trigger-phrase rule is as follows: “It is noteworthy that the patient had received 0.5 mg Ativan x2 and morphine earlier in the afternoon and there is a concern that this may have contributed to his altered mental status.” In this context sentence, an opioid drug (“morphine”) is identified alongside a trigger phrase (“contributed to”). The Spark-NLP NER model identified the AE term as *altered mental status*. This term was resolved to the Medical Dictionary for Regulatory Activities (MedDRA) term *mental state abnormal* using Usagi (Observational Health Data Sciences and Informatics) and the corresponding UMLS concept, as in the section on disambiguation of the ORADEs below. The candidate ORADE pair generated from this information is *morphine-mental state abnormal*.

Second, the antidote-based ADE detection rule: We identified ORADE context sentences by identifying mentions of the drug naloxone, an FDA-approved medication that reverses an overdose caused by an opioid drug. To capture mentions of naloxone and opioid drugs that did not co-occur in the same sentence, we searched through the preceding and following 3 sentences. Antidote signals have been used in detecting ADRs in published literature reports [25,26].

An example of an antidote-based ADE detection rule is as follows: “He received dilaudid q 2 hr at 7:30 am, 9:30 am, 11:30 am. Code blue was called for respiratory arrest (unwitnessed). 0.4 mg of Narcan IV was administered followed by 1 mg of IV Narcan. This resulted in improvement of his respiratory status and regain of his consciousness.” In this example, the antidote-based detection rule captures mentions of naloxone in the context sentence, respiratory arrest in the preceding sentence, and the Dilaudid mention in the prior sentence to generate the candidate ORADE pair *Dilaudid-respiratory arrest*.

NER to Identify ORADEs in Clinical Text

Having detected opioid drug terms, we used a pretrained NER model, Spark NLP for Healthcare, which uses deep learning-based NER to identify possible AE terms in sentences. The model is a biLSTM, convolutional neural network, character-based deep learning model trained using biomedical NER data sets such as AnatEM, BC5CDR, BC4CHEMD, BioNLP13CG, JNLPBA, Linnaeus, NCBI-Disease, and S800 [27]. After identifying the AE terms in the context sentences, we connected all opioid mentions in the context sentences with the AE terms to create candidate ORADE pairs.

Disambiguation of ORADEs

AE terms can appear with different spellings, spelling errors, or abbreviations; therefore, we used the UMLS to map the free text to standardized concepts. We used ScispaCy to map the raw phrase found in the discharge summary to the standard UMLS translation of the concept [28]. Furthermore, we used Usagi to obtain the MedDRA term for the UMLS concept. The identified MedDRA AE term is mapped to the opioid drug term to create a candidate ORADE pair that incorporates standardized MedDRA terminology, including preferred terms (PTs) or lower-level terms (LLTs).

Prototype User Testing and Feedback From Participants

We recruited 15 CDER staff members to assess the various features, functionalities, and graphic visualizations. They were

experienced in the use of web-based software tools but were not involved in the development of this prototype.

Testing Design and Conduct

A testing guide was provided that included login instructions, descriptions and screenshots of the application features and components, and instructions for exporting outputs. Test participants worked remotely, were not monitored, and were given 1 week to complete their testing. Participants were free to explore the application for their regulatory work.

For user testing, we extracted from the MIMIC III database a subset of discharge summaries filtered for an opioid drug keyword. The subset included 31,052 notes corresponding to 30,326 hospital admission events for 24,539 patients.

Metrics

Each participant completed an anonymous electronic survey covering technical operation, ease of navigating and interpreting various visualizations, and user satisfaction for drug safety and research (Multimedia Appendix 3).

Results

ORADE Detection

The prototype application successfully detected ORADEs that correspond to known opioid drug toxicities. The most commonly identified opioid drugs and the top 3 most frequent ORADEs per drug are summarized in Table 1.

To assess the contribution of keywords with trigger phrases and antidote (naloxone) signals for ORADE detection, we examined quantitative parameters for a filtered MIMIC III data subset, as shown in Table 2.

Table 2 shows that keywords with trigger phrases detect most unique AEs and candidate ORADEs in context discharge summaries. In comparison, the approach based on the antidote (ie, naloxone) makes a much smaller relative contribution to ORADE detection.

Table 1. Opioid-related adverse drug event detection from the text of the electronic health record discharge summaries.

Opioid drug class	Most frequently identified opioid drug	Top 3 most frequently identified opioid-related adverse drug events
Natural	Morphine	Hypotension; somnolence; nausea
Semisynthetic	Hydromorphone	Confusion; hypotension; agitation
Synthetic	Fentanyl	Hypotension; adverse reaction; hepatitis C

Table 2. Relative contribution of keywords with trigger phrase and antidote (ie, naloxone) signals for candidate opioid-related adverse drug event detection. International Classification of Diseases (Ninth Revision) code E935.2, which specifies opioids and other narcotics causing adverse effects in therapeutic use, was used to create a filtered subset of Medical Information Mart for Intensive Care III (MIMIC III) discharge summaries having at least one opioid-related adverse drug event pair.

	ORADE ^a detection based on keywords with trigger phrases	ORADE detection based only on antidote (naloxone) signals	ORADE detection based on both trigger phrases and antidote signals
Number of unique opioid drugs detected (n=12)	12 (100%)	6 (50%)	6 (50%)
Number of unique AEs ^b detected (n=117)	110 (94%)	8 (7%)	15 (13%)
Number of unique candidate ORADE pairs (n=219)	205 (94%)	13 (6%)	17 (8%)
Number of discharge summaries (n=101)	95 (94%)	8 (8%)	12 (12%)
Number of unique patients (n=101)	94 (94%)	8 (8%)	12 (12%)

^aORADE: opioid-related adverse drug event.

^bAE: adverse event.

Error Analysis

An error analysis was performed to characterize incorrect candidate ORADE pairs and is summarized with mitigation strategies in [Table 3](#).

Table 3. Error analysis of false-positive and false-negative candidate opioid-related adverse drug event pairs.

Category/type and relative frequency	Example	Mitigation strategy
False positive		
Drug indication pairs ^a	Text: “She was given fentanyl for the back pain with subsequent hypotension.” Incorrect candidate ORADE ^b pair: <i>fentanyl-back pain</i>	Condition terms that include “pain” are excluded.
Drug/medication change events ^c	Text: “She was changed from Percocet to Ultram due to nausea, which resolved.” Incorrect candidate ORADE pair: <i>Ultram-nausea</i>	The context sentence is scanned for the following phrases using regular expressions: “change to,” “switch to,” “change from drug X to drug Y,” or “switch from drug X to drug Y.” Candidate opioid drug-drug medication change event pairs so generated are excluded.
Negated ADE ^d mentions where the AE ^e is not due to a drug ^f	Text: “No further apneic events.” Incorrect candidate ADE: <i>apneic events</i>	The assertion module in Spark NLP ^g for Healthcare is used to detect negation so that any negated condition term is not included in a candidate ORADE pair.
False negative		
Concept fragmentation ^c	Text: “She had been treated with high dose fentanyl and benzodiazepines which were the most likely cause of delirium.... She was also found to be severely constipated. # Constipation: patient developed severe constipation related to pain medication. She was manually disimpacted and started on an aggressive [sic] bowel regimen.” Missed candidate ORADE pair: <i>opioid drug-constipation</i>	<i>Severe constipation</i> was detected, but the current model could not find which pain medication it was related to. To resolve, we will explore more data and consider other rules or models.
Entity not recognized as an AE	Text: “He does endorse decreased sleep latency, falling asleep in less than 5 minutes, and also questionable daytime hypersomnolence, but denies morning headaches. Of note, patient received prescription for Vicodin upon discharge from ED on [**2173-8-28**].” Missed candidate AE: <i>hypersomnolence</i>	To resolve, we will explore more data and consider other rules or models.
Entity not recognized as an opioid drug ^c	Text: “His hospital course was complicated by a respiratory code on the floor attributed to respiratory suppression from narcotics.” Missed candidate drug: <i>narcotics</i>	<i>Narcotics</i> could be added to the opioid keyword list. To resolve to a specific opioid drug, we will explore more data and consider other rules or models.

^aMost commonly encountered error.

^bORADE: opioid-related adverse drug event.

^cModerately encountered error.

^dADE: adverse drug event.

^eAE: adverse event.

^fRarely encountered error.

^gNLP: natural language processing.

Prototype Application Performance Metrics

We calculated the performance metrics accuracy, recall, precision, and F_1 -score using conventional mathematical formulas [14]. The AI-enabled model achieved accuracy, recall, precision, and F_1 -scores of 0.66, 0.69, 0.64, and 0.67, respectively, based on the test subset of 46 discharge summaries. Candidate ORADE pairs generated with this software prototype are hypothetical and do not indicate causality or absolute risk for an association. Further assessment is required by subject matter experts.

Prototype Application Analytics Dashboard

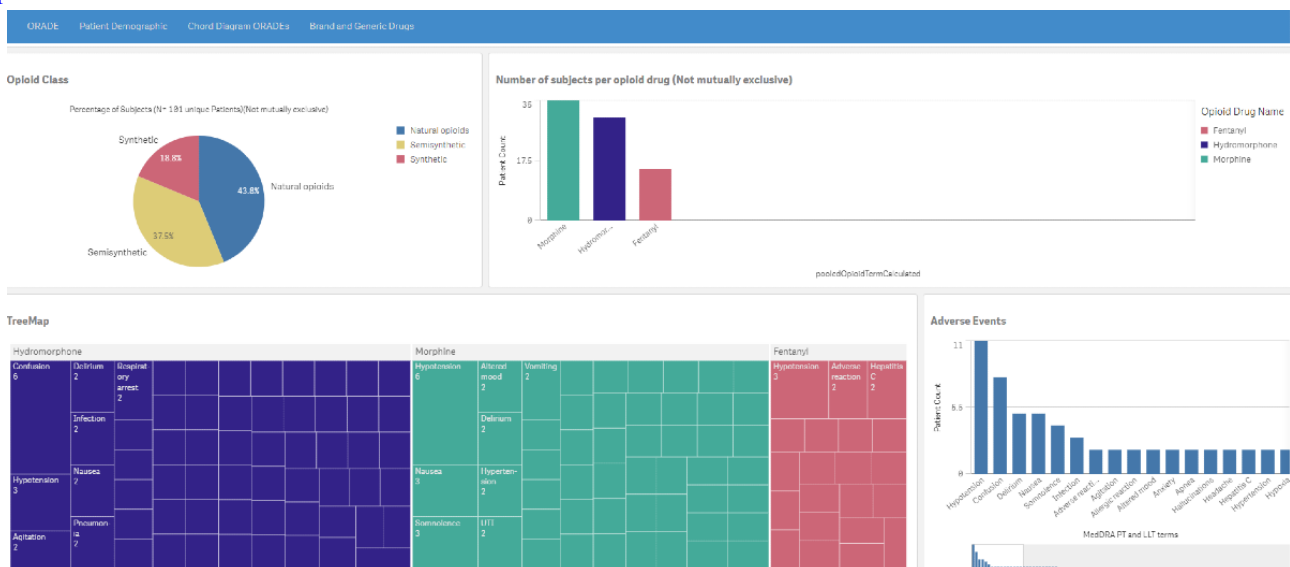
The Qlik Sense data analytics platform (QlikTech International AB) was used to implement the SPINEL dashboard with interactive graphics, visualizations, and line listings. The landing page (Figure 2) has 4 sheet tabs: ORADE, Patient Demographic, Chord Diagram ORADEs, and Brand and Generic Drugs. They are described below with morphine used as an arbitrarily selected opioid drug for the graphics and visualizations.

The ORADE tab (Figure 3) has four components: (1) a pie chart that shows subsets of the 3 classes of opioid drugs, (2) a histogram of all subjects per drug, (3) a tree map of the MedDRA PTs and LLTs for each drug, and (4) a second histogram of patient count by MedDRA PT and LLT for the selected drug(s) of interest.

Figure 2. The landing page for SPINEL (Supporting Pharmacovigilance by Leveraging Artificial Intelligence Methods to Analyze Electronic Health Records Data) depicting a pie chart (upper left) of the 3 opioid classes; a histogram (upper right) of the subject counts per opioid drug; a tree map (lower left) of the electronic health record–derived opioid-related adverse drug profiles, where the adverse events for each opioid drug are represented by nested rectangles and the size of the nested rectangle relates to the patient count per adverse event; and a histogram (lower right) of patient count by MedDRA (Medical Dictionary for Regulatory Activities) preferred term and lower-level term for the drugs. A higher resolution version of this figure is available in [Multimedia Appendix 1](#).



Figure 3. The opioid-related adverse drug page depicting a pie chart (upper left) and a histogram (upper right) of the 101 subjects who received at least one opioid class drug, a tree map (lower left) of the electronic health record–derived opioid-related adverse drug profile for the most frequently identified opioid class drugs, and a histogram (lower right) of patient count by MedDRA (Medical Dictionary for Regulatory Activities) preferred term and lower-level term for the top 3 most frequently identified opioid-related adverse drugs. A higher resolution version of this figure is available in [Multimedia Appendix 1](#).



The patient demographic tab (Figure 4) includes the following components: (1) a histogram of age, (2) a pie chart of gender, (3) another histogram of ethnicity, and (4) a line listing of the individual patients with AEs and associated demographics.

The chord diagram tab (Figure 5) displays a graphic to visually explore interconnections between opioid drugs and AE mentions.

The brand and generic drugs tab (Figure 6) includes multiple displays: (1) a pie chart with the percentage patient count by brand or generic drug type, (2) a stacked bar chart of patients by opioid class and drug type, and (3) a searchable, scrollable spreadsheet listing of the drug name, drug type, and adverse events associated with the subject IDs.

Figure 4. Patient demographics page depicting a histogram for morphine treated-patients by age (upper left), a pie chart for gender (upper right), a histogram for ethnicity (lower left), and a line listing (lower right) of the individual patients with adverse events and associated demographics. Morphine is an arbitrarily selected natural opioid drug. A higher resolution version of this figure is available in [Multimedia Appendix 1](#).

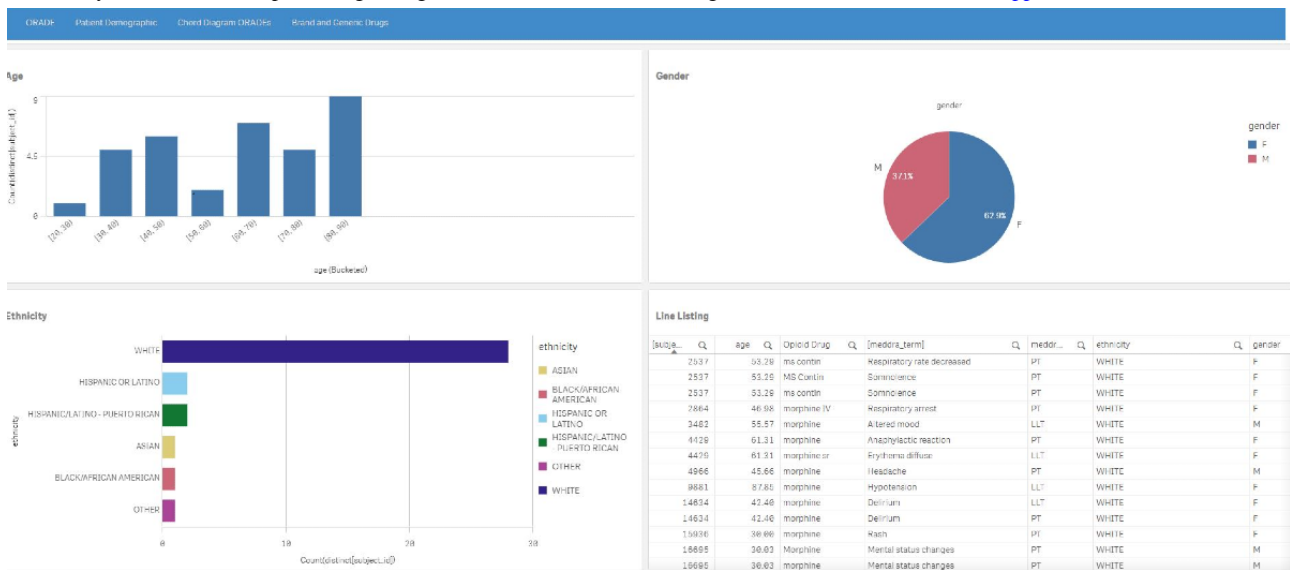


Figure 5. Cord diagram page visually depicting the interconnections between the opioid drug of interest (morphine in this example) and adverse event mentions as derived from the electronic health record discharge summaries. The larger the caliber of the connecting cord, the higher the adverse drug event frequency. Morphine is an arbitrarily selected natural opioid drug. A higher resolution version of this figure is available in [Multimedia Appendix 1](#).

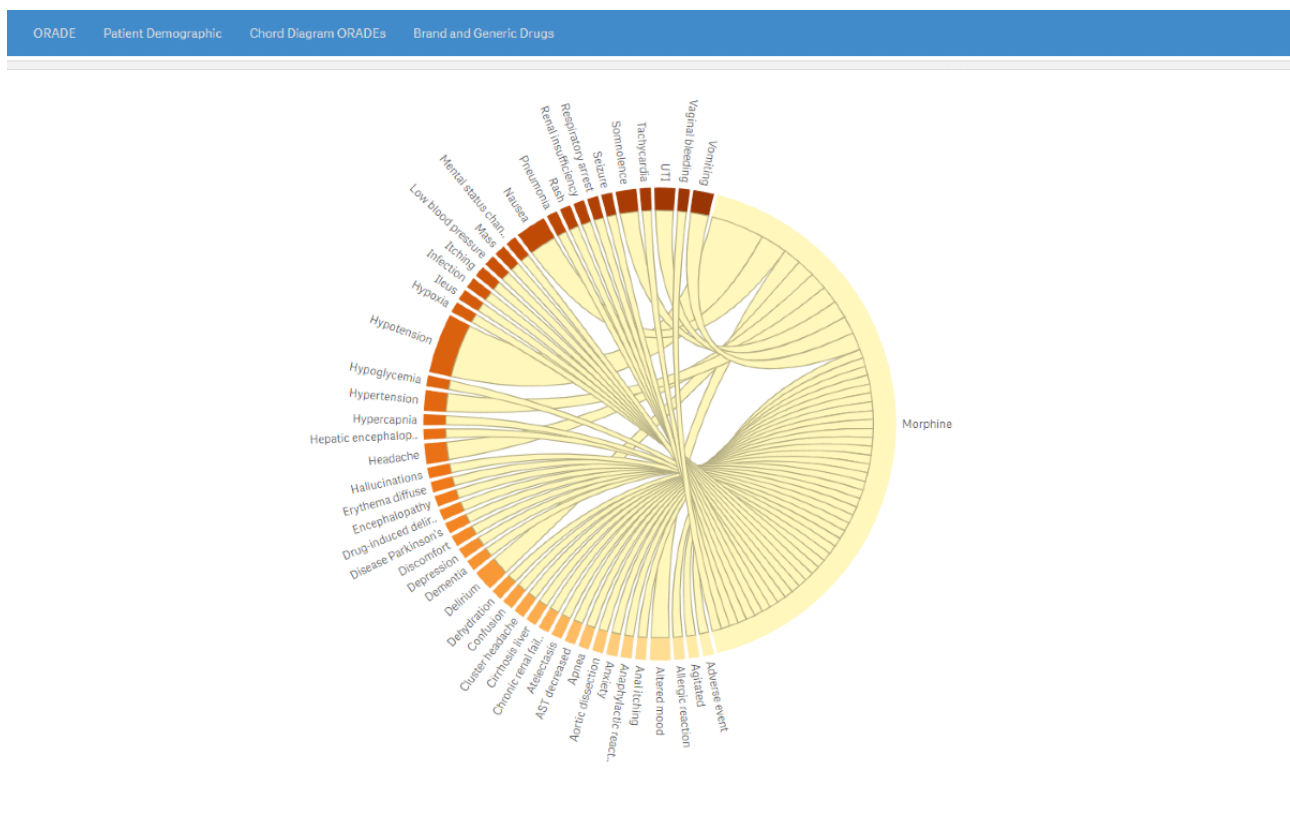
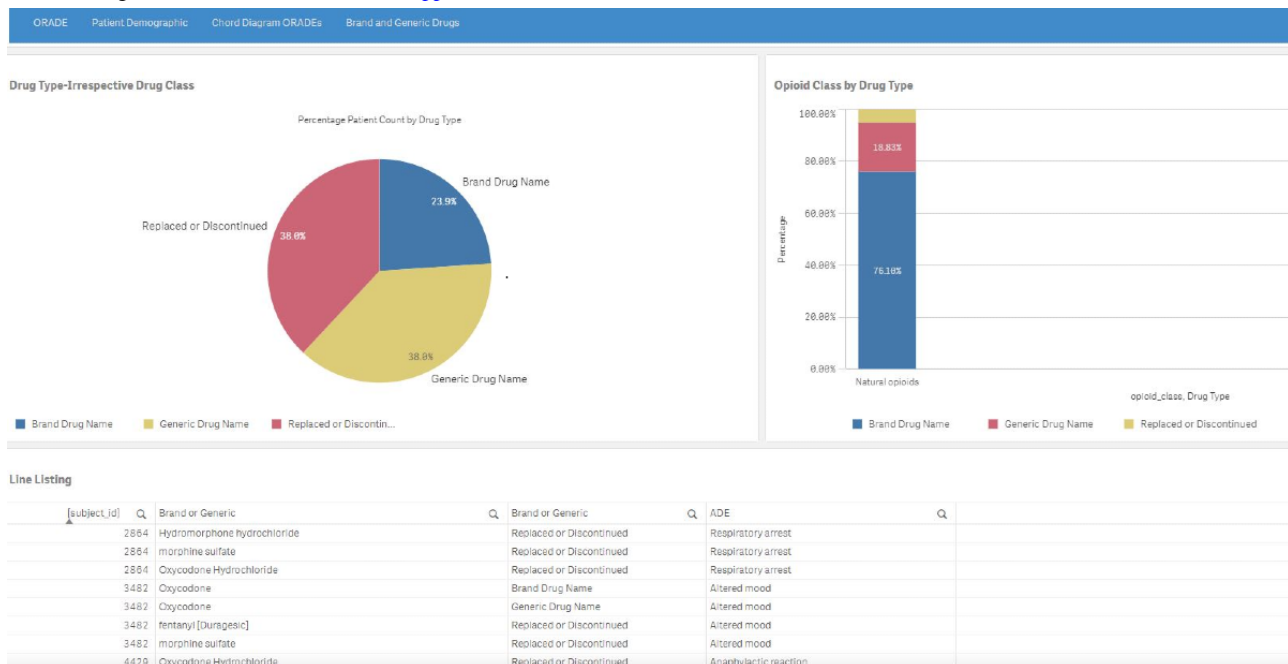


Figure 6. Brand and generic drugs page depicting a pie chart (upper left) of brand, generic, or replaced/discontinued drug type, a stacked bar chart (upper right) of patients by type, and a line listing (lower section) of the patients by drug name, drug type, and adverse events. A higher resolution version of this figure is available in [Multimedia Appendix 1](#).



Results of User Testing

SPINEL was assessed as a highly desirable prototype that satisfies end user needs for supporting opioid drug safety signal detection from EHR data. The application was easy to use, the visualizations enhanced detection of drug safety signals, and the prototype ranked high in saving time compared to manual chart review. Survey results were based on a Likert rating scale ([Multimedia Appendix 4](#)).

Fifteen FDA staff completed the survey questionnaire with 11 providing observational feedback. Participant feedback uncovered a few minor bugs and indicated the following areas for potential improvement: (1) enlarge the tabs and visualizations, (2) enable more flexibility and customizations to fit each end user's needs, (3) provide additional instructional resources to enhance learning about the various features and functionalities, (4) add multiple graph export functionality, and (5) add project summaries. Possible mitigation strategies include adding a slider bar with zoom function for the more complex visualizations, providing an instructional video on the application's features and functionalities, providing tool-tip pop-ups and a supplemental "user tips" guide to highlight key features or functionality, modifying the export function to accommodate multiple graphics, and developing a customizable user portal to include project summaries.

Discussion

Principal Results

The AI-enabled SPINEL prototype successfully detects known opioid drug toxicities from free text in EHRs and provides a framework to uncover emerging safety data that could potentially augment regulatory review and decision-making. Automated processing and analysis of EHR data reduces the

research burden compared to manual chart review, saving considerable time and effort. The prototype expedites the quick perusal of data for trends and patterns reflecting drug toxicities while facilitating drilling down into the data to patient-level line listing information. FDA participants conveyed high satisfaction ratings for this prototype and acknowledged its potential to add value in harnessing unstructured text in EHRs for pharmacovigilance.

In applying our AI-based model, we limited our analysis to discharge summaries because published studies confirm that discharge summaries are the best subsection of the EHR for gathering information about ADEs reported by physicians [29-31]. In reviewing the discharge summaries, we observed considerable heterogeneity in the quality of reporting and the depth of detail conveyed about possible ORADEs, which could affect the accuracy and other performance metrics for the software application. We applied 2 rule-based algorithms to enhance ORADE detection from discharge summaries. Our results demonstrate that the majority of candidate ORADE pairs and context discharge summaries are detected using keywords with trigger phrases. As described in published literature [32], this approach to searching for drug safety signals is best for uncovering ADEs potentially related to specific drug products as delineated in the keyword list (opioids in our use case). As new drug products are approved by the FDA, the keyword list would need manual updating to keep it current. However, for broader and more generalized searching, this could become cumbersome, as new keyword lists would need to be manually compiled for each drug grouping or class of interest.

The accuracy, recall, and precision of this prototype will need to be improved to better align with established NLP processors. Two steps to be considered in future work to improve performance are (1) leveraging information from established drug databases, such as the DailyMed database of the most

recent FDA-approved drug product labels to filter out false positive ORADEs due to drug-indication pairs and (2) using large language models (LLMs) such as GPT-4 [33], BioGPT [34], or GatorTron [35] to improve capture of mentions of opioid drugs and ADE terms that may be separated by multiple paragraphs.

Limitations

This project encountered three main challenges and limitations. First, patient cohort identification: Use of ICD-9 codes to prescreen discharge summaries for potential cases of ORADEs could be impacted by selection and misclassification biases resulting in a subset that may not reflect the total number of ORADE cases in the MIMIC III data. These biases could result in a skewed patient sample wherein there may be missed patients with ORADEs or patients incorrectly classified as having an ORADE due to erroneous coding. In addition, in focusing only on the free-text discharge summaries, we may have missed patients whose ORADEs were captured only in other text reports that we did not explore, such as physician notes, nursing notes, and consultation reports. Together, these issues may prevent us from capturing the full extent and scope of patients experiencing ORADEs from the MIMIC III EHRs. In future work, a more robust approach to identifying patients with ORADEs will be considered, including use of a standardized annotation guideline and reporting of interannotator agreement scores related to development of a reference data set; possible inclusion of objective components for case ascertainment, such as laboratory or medical imaging abnormalities; and expanding the scope of reports assessed to include physician notes, nursing notes, and consultation reports, where available, in addition to discharge summaries. Second was the use of MIMIC III. The single-center MIMIC III EHR database may not reflect the broad diversity of the US population, which could limit generalizability for

drug safety surveillance to larger and more diversified domains and lend to potentially biased assessments. Third, the lack of a publicly available reference standard data set hindered efforts to evaluate the NLP component of our AI-enabled model in detecting ADE safety signals from text in EHRs. The small size of our reference data set risked overfitting and biased assessments.

There were limitations inherent in the user testing procedures. User testing was unmonitored and conducted without prespecified tasks. This approach accommodated participants working in remote locations to explore the software in their regulatory work. However, direct observation by a facilitator may have enabled us to gather more details about end-user experience. Additionally, the sample size of intended users was small. Feedback from a larger group of CDER regulatory staff may be more informative about the potential impact on their regulatory work and decision-making.

Conclusions

SPINEL, our novel AI-enabled software, extracts ORADEs from free-text discharge summaries in EHRs, streamlines workflow, and augments access to real world data for pharmacovigilance. Detecting opioid safety signals from EHRs enhances the capacity to harness an important yet underutilized resource of clinically relevant information for regulatory review and decision-making.

Future work will explore detecting newly emerging opioid drug safety issues using a larger and more diversified EHR database, investigating various methods to improve NLP performance, resolving application features per FDA participant feedback, and integrating knowledge graphs to interconnect information from EHRs with reports published in the literature.

Acknowledgments

This project was supported in part by appointment to the research participation program at the Center for Drug Evaluation and Research (CDER), which is administered by the Oak Ridge Institute for Science and Education for the US Food and Drug Administration (FDA) and by the Lister Hill National Center for Biomedical Communications of the National Library of Medicine of the National Institutes of Health. Funding support was received from the FDA, CDER, and the Opioid Research Program. We also acknowledge Henry Francis, MD, and Mahmudul Hasan, PhD, for their contributions to the project.

Disclaimers

This publication reflects the views of the authors and should not be construed to represent US Food and Drug Administration views or policies. The mention of commercial products, their sources, or their use in connection with material reported herein is not to be construed as either an actual or implied endorsement of such products by the US Department of Health and Human Services.

Authors' Contributions

AS and RH contributed to project concept and design; SAH and RH contributed to data processing; AS, AV, and KZ contributed to the reference standard; SAH and RH contributed to the natural language processing and artificial intelligence-enabled models; RH contributed to dashboard configuration; AS, RH, SAH, RJ, AH, AV, KZ, and AR contributed to interpretation of the results; AS and SAH contributed to drafting of the manuscript; AS, RH, AR, RJ, and MA contributed to manuscript revision; and AS, RH, AR, RJ, and MA contributed to final approval of the manuscript.

Conflicts of Interest

None declared.

Multimedia Appendix 1

High resolution versions of Figures 1-6.

[[PDF File \(Adobe PDF File\), 1379 KB - ai_v2i1e45000_app1.pdf](#)]

Multimedia Appendix 2

List of 58 newly curated trigger phrases.

[[PDF File \(Adobe PDF File\), 44 KB - ai_v2i1e45000_app2.pdf](#)]

Multimedia Appendix 3

Anonymous user survey questions.

[[PDF File \(Adobe PDF File\), 81 KB - ai_v2i1e45000_app3.pdf](#)]

Multimedia Appendix 4

User Survey: Likert scale ratings from user testing.

[[PDF File \(Adobe PDF File\), 35 KB - ai_v2i1e45000_app4.pdf](#)]

References

1. Alatawi YM, Hansen RA. Empirical estimation of under-reporting in the U.S. Food and Drug Administration Adverse Event Reporting System (FAERS). *Expert Opin Drug Saf* 2017 Jul;16(7):761-767. [doi: [10.1080/14740338.2017.1323867](https://doi.org/10.1080/14740338.2017.1323867)] [Medline: [28447485](https://pubmed.ncbi.nlm.nih.gov/28447485/)]
2. La Grenade L, Graham DJ, Nourjah P. Underreporting of hemorrhagic stroke associated with phenylpropanolamine. *JAMA* 2001 Dec 26;286(24):3081. [doi: [10.1001/jama.286.24.3081](https://doi.org/10.1001/jama.286.24.3081)] [Medline: [11754672](https://pubmed.ncbi.nlm.nih.gov/11754672/)]
3. Harpaz R, DuMouchel W, Schuemie M, Bodenreider O, Friedman C, Horvitz E, et al. Toward multimodal signal detection of adverse drug reactions. *J Biomed Inform* 2017 Dec;76:41-49 [FREE Full text] [doi: [10.1016/j.jbi.2017.10.013](https://doi.org/10.1016/j.jbi.2017.10.013)] [Medline: [29081385](https://pubmed.ncbi.nlm.nih.gov/29081385/)]
4. Coloma PM, Trifirò G, Patadia V, Sturkenboom M. Postmarketing safety surveillance: where does signal detection using electronic healthcare records fit into the big picture? *Drug Saf* 2013 Mar;36(3):183-197. [doi: [10.1007/s40264-013-0018-x](https://doi.org/10.1007/s40264-013-0018-x)] [Medline: [23377696](https://pubmed.ncbi.nlm.nih.gov/23377696/)]
5. Spence MM, Nguyen LM, Hui RL, Chan J. Evaluation of clinical and safety outcomes associated with conversion from brand-name to generic tacrolimus in transplant recipients enrolled in an integrated health care system. *Pharmacotherapy* 2012 Nov;32(11):981-987. [doi: [10.1002/phar.1130](https://doi.org/10.1002/phar.1130)] [Medline: [23074134](https://pubmed.ncbi.nlm.nih.gov/23074134/)]
6. Nie X, Yu Y, Jia L, Zhao H, Chen Z, Zhang L, et al. Signal detection of pediatric drug-induced coagulopathy using routine electronic health records. *Front Pharmacol* 2022;13:935627 [FREE Full text] [doi: [10.3389/fphar.2022.935627](https://doi.org/10.3389/fphar.2022.935627)] [Medline: [35935826](https://pubmed.ncbi.nlm.nih.gov/35935826/)]
7. Harpaz R, Vilar S, Dumouchel W, Salmasian H, Haerian K, Shah NH, et al. Combining signals from spontaneous reports and electronic health records for detection of adverse drug reactions. *J Am Med Inform Assoc* 2013 May 01;20(3):413-419 [FREE Full text] [doi: [10.1136/amiajnl-2012-000930](https://doi.org/10.1136/amiajnl-2012-000930)] [Medline: [23118093](https://pubmed.ncbi.nlm.nih.gov/23118093/)]
8. Ford E, Carroll JA, Smith HE, Scott D, Cassell JA. Extracting information from the text of electronic medical records to improve case detection: a systematic review. *J Am Med Inform Assoc* 2016 Sep;23(5):1007-1015 [FREE Full text] [doi: [10.1093/jamia/ocv180](https://doi.org/10.1093/jamia/ocv180)] [Medline: [26911811](https://pubmed.ncbi.nlm.nih.gov/26911811/)]
9. Murff HJ, Forster AJ, Peterson JF, Fiskio JM, Heiman HL, Bates DW. Electronically screening discharge summaries for adverse medical events. *J Am Med Inform Assoc* 2003;10(4):339-350 [FREE Full text] [doi: [10.1197/jamia.M1201](https://doi.org/10.1197/jamia.M1201)] [Medline: [12668691](https://pubmed.ncbi.nlm.nih.gov/12668691/)]
10. Cantor MN, Feldman HJ, Triola MM. Using trigger phrases to detect adverse drug reactions in ambulatory care notes. *Qual Saf Health Care* 2007 Apr;16(2):132-134 [FREE Full text] [doi: [10.1136/qshc.2006.020073](https://doi.org/10.1136/qshc.2006.020073)] [Medline: [17403760](https://pubmed.ncbi.nlm.nih.gov/17403760/)]
11. Friedman C, Shagina L, Lussier Y, Hripesak G. Automated encoding of clinical documents based on natural language processing. *J Am Med Inform Assoc* 2004;11(5):392-402 [FREE Full text] [doi: [10.1197/jamia.M1552](https://doi.org/10.1197/jamia.M1552)] [Medline: [15187068](https://pubmed.ncbi.nlm.nih.gov/15187068/)]
12. Aronson AR, Lang F. An overview of MetaMap: historical perspective and recent advances. *J Am Med Inform Assoc* 2010;17(3):229-236 [FREE Full text] [doi: [10.1136/jamia.2009.002733](https://doi.org/10.1136/jamia.2009.002733)] [Medline: [20442139](https://pubmed.ncbi.nlm.nih.gov/20442139/)]
13. Savova GK, Masanz JJ, Ogren PV, Zheng J, Sohn S, Kipper-Schuler KC, et al. Mayo clinical Text Analysis and Knowledge Extraction System (cTAKES): architecture, component evaluation and applications. *J Am Med Inform Assoc* 2010;17(5):507-513 [FREE Full text] [doi: [10.1136/jamia.2009.001560](https://doi.org/10.1136/jamia.2009.001560)] [Medline: [20819853](https://pubmed.ncbi.nlm.nih.gov/20819853/)]

14. Murphy RM, Klopotoska JE, de Keizer NF, Jager KJ, Leopold JH, Dongelmans DA, et al. Adverse drug event detection using natural language processing: A scoping review of supervised learning methods. *PLoS One* 2023;18(1):e0279842 [FREE Full text] [doi: [10.1371/journal.pone.0279842](https://doi.org/10.1371/journal.pone.0279842)] [Medline: [36595517](https://pubmed.ncbi.nlm.nih.gov/36595517/)]
15. Jagannatha A, Liu F, Liu W, Yu H. Overview of the first natural language processing challenge for extracting medication, indication, and adverse drug events from electronic health record notes (MADE 1.0). *Drug Saf* 2019 Jan;42(1):99-111 [FREE Full text] [doi: [10.1007/s40264-018-0762-z](https://doi.org/10.1007/s40264-018-0762-z)] [Medline: [30649735](https://pubmed.ncbi.nlm.nih.gov/30649735/)]
16. Henry S, Buchan K, Filannino M, Stubbs A, Uzuner O. 2018 n2c2 shared task on adverse drug events and medication extraction in electronic health records. *J Am Med Inform Assoc* 2020 Jan 01;27(1):3-12 [FREE Full text] [doi: [10.1093/jamia/ocz166](https://doi.org/10.1093/jamia/ocz166)] [Medline: [31584655](https://pubmed.ncbi.nlm.nih.gov/31584655/)]
17. McKnight W, Dolezal J. Healthcare natural language processing. GigaOm. URL: <https://research.gigaom.com/report/healthcare-natural-language-processing/> [accessed 2023-03-21]
18. Davies EC, Green CF, Taylor S, Williamson PR, Mottram DR, Pirmohamed M. Adverse drug reactions in hospital in-patients: a prospective analysis of 3695 patient-episodes. *PLoS One* 2009;4(2):e4439 [FREE Full text] [doi: [10.1371/journal.pone.0004439](https://doi.org/10.1371/journal.pone.0004439)] [Medline: [19209224](https://pubmed.ncbi.nlm.nih.gov/19209224/)]
19. Urman RD, Seger DL, Fiskio JM, Neville BA, Harry EM, Weiner SG, et al. The burden of opioid-related adverse drug events on hospitalized previously opioid-free surgical patients. *J Patient Saf* 2021 Mar 01;17(2):e76-e83. [doi: [10.1097/PTS.0000000000000566](https://doi.org/10.1097/PTS.0000000000000566)] [Medline: [30672762](https://pubmed.ncbi.nlm.nih.gov/30672762/)]
20. Johnson AEW, Pollard TJ, Shen L, Lehman LH, Feng M, Ghassemi M, et al. MIMIC-III, a freely accessible critical care database. *Sci Data* 2016 May 24;3:160035 [FREE Full text] [doi: [10.1038/sdata.2016.35](https://doi.org/10.1038/sdata.2016.35)] [Medline: [27219127](https://pubmed.ncbi.nlm.nih.gov/27219127/)]
21. Hougland P, Xu W, Pickard S, Masheter C, Williams SD. Performance of International Classification Of Diseases, 9th Revision, Clinical Modification codes as an adverse drug event surveillance system. *Med Care* 2006 Jul;44(7):629-636. [doi: [10.1097/01.mlr.0000215859.06051.77](https://doi.org/10.1097/01.mlr.0000215859.06051.77)] [Medline: [16799357](https://pubmed.ncbi.nlm.nih.gov/16799357/)]
22. Eyre H, Chapman AB, Peterson KS, Shi J, Alba PR, Jones MM, et al. Launching into clinical space with medspaCy: a new clinical text processing toolkit in Python. *AMIA Annu Symp Proc* 2021;2021:438-447 [FREE Full text] [Medline: [35308962](https://pubmed.ncbi.nlm.nih.gov/35308962/)]
23. Tang Y, Yang J, Ang PS, Dorajoo SR, Foo B, Soh S, et al. Detecting adverse drug reactions in discharge summaries of electronic medical records using Readpeer. *Int J Med Inform* 2019 Aug;128:62-70. [doi: [10.1016/j.ijmedinf.2019.04.017](https://doi.org/10.1016/j.ijmedinf.2019.04.017)] [Medline: [31160013](https://pubmed.ncbi.nlm.nih.gov/31160013/)]
24. Harkema H, Dowling JN, Thornblade T, Chapman WW. ConText: an algorithm for determining negation, experienter, and temporal status from clinical reports. *J Biomed Inform* 2009 Oct;42(5):839-851 [FREE Full text] [doi: [10.1016/j.jbi.2009.05.002](https://doi.org/10.1016/j.jbi.2009.05.002)] [Medline: [19435614](https://pubmed.ncbi.nlm.nih.gov/19435614/)]
25. Handler SM, Hanlon JT, Perera S, Roumani YF, Nace DA, Fridsma DB, et al. Consensus list of signals to detect potential adverse drug reactions in nursing homes. *J Am Geriatr Soc* 2008 May;56(5):808-815 [FREE Full text] [doi: [10.1111/j.1532-5415.2008.01665.x](https://doi.org/10.1111/j.1532-5415.2008.01665.x)] [Medline: [18363678](https://pubmed.ncbi.nlm.nih.gov/18363678/)]
26. Kane-Gill SL, Bellamy CJ, Verrico MM, Handler SM, Weber RJ. Evaluating the positive predictive values of antidote signals to detect potential adverse drug reactions (ADRs) in the medical intensive care unit (ICU). *Pharmacoepidemiol Drug Saf* 2009 Dec;18(12):1185-1191. [doi: [10.1002/pds.1837](https://doi.org/10.1002/pds.1837)] [Medline: [19728294](https://pubmed.ncbi.nlm.nih.gov/19728294/)]
27. Kocaman V, Talby D. Accurate Clinical and Biomedical Named Entity Recognition at Scale. *Software Impacts* 2022 Aug;13:100373 [FREE Full text] [doi: [10.1016/j.simpa.2022.100373](https://doi.org/10.1016/j.simpa.2022.100373)]
28. Neumann M, King D, Beltagy I. ScispaCy: fast and robust models for biomedical natural language processing. In: *Proceedings of the 18th BioNLP Workshop and Shared Task*. 2019 Presented at: 18th BioNLP Workshop and Shared Task; Aug 1, 2019; Florence, Italy p. 319-327. [doi: [10.18653/v1/W19-5034](https://doi.org/10.18653/v1/W19-5034)]
29. Tinoco A, Evans RS, Staes CJ, Lloyd JF, Rothschild JM, Haug PJ. Comparison of computerized surveillance and manual chart review for adverse events. *J Am Med Inform Assoc* 2011;18(4):491-497 [FREE Full text] [doi: [10.1136/amiajnl-2011-000187](https://doi.org/10.1136/amiajnl-2011-000187)] [Medline: [21672911](https://pubmed.ncbi.nlm.nih.gov/21672911/)]
30. Anthes AM, Harinstein LM, Smithburger PL, Seybert AL, Kane-Gill SL. Improving adverse drug event detection in critically ill patients through screening intensive care unit transfer summaries. *Pharmacoepidemiol Drug Saf* 2013 May;22(5):510-516. [doi: [10.1002/pds.3422](https://doi.org/10.1002/pds.3422)] [Medline: [23440931](https://pubmed.ncbi.nlm.nih.gov/23440931/)]
31. Aguilera C, Agustí A, Pérez E, Gracia RM, Diogène E, Danés I. Spontaneously reported adverse drug reactions and their description in hospital discharge reports: A retrospective study. *J Clin Med* 2021 Jul 26;10(15):3293 [FREE Full text] [doi: [10.3390/jcm10153293](https://doi.org/10.3390/jcm10153293)] [Medline: [34362076](https://pubmed.ncbi.nlm.nih.gov/34362076/)]
32. Luo Y, Thompson WK, Herr TM, Zeng Z, Berendsen MA, Jonnalagadda SR, et al. Natural language processing for EHR-based pharmacovigilance: A structured review. *Drug Saf* 2017 Nov;40(11):1075-1089. [doi: [10.1007/s40264-017-0558-6](https://doi.org/10.1007/s40264-017-0558-6)] [Medline: [28643174](https://pubmed.ncbi.nlm.nih.gov/28643174/)]
33. GPT-4 technical report. OpenAI. URL: <https://cdn.openai.com/papers/gpt-4.pdf> [accessed 2023-06-19]
34. Luo R, Sun L, Xia Y, Qin T, Zhang S, Poon H, et al. BioGPT: generative pre-trained transformer for biomedical text generation and mining. *Brief Bioinform* 2022 Nov 19;23(6):bbac409. [doi: [10.1093/bib/bbac409](https://doi.org/10.1093/bib/bbac409)] [Medline: [36156661](https://pubmed.ncbi.nlm.nih.gov/36156661/)]
35. Yang X, Chen A, PourNejatian N, Shin HC, Smith KE, Parisien C, et al. A large language model for electronic health records. *NPJ Digit Med* 2022 Dec 26;5(1):194 [FREE Full text] [doi: [10.1038/s41746-022-00742-2](https://doi.org/10.1038/s41746-022-00742-2)] [Medline: [36572766](https://pubmed.ncbi.nlm.nih.gov/36572766/)]

Abbreviations

ADE: adverse drug event
ADR: adverse drug reaction
AE: adverse event
AI: artificial intelligence
BERT: bidirectional encoder representations from transformers
CDER: Center for Drug Evaluation and Research
CNN: convolutional neural network
CRF: conditional random field
cTAKES: Clinical Text Analysis and Knowledge Extraction System
EHR: electronic health record
FAERS: FDA Adverse Events Reporting System
FDA: US Food and Drug Administration
GPT: generative pretrained transformer
ICD-9: International Classification of Diseases, Ninth Revision
LLM: large language model
LLT: lower-level term
LSTM: long short-term memory
MedDRA: Medical Dictionary for Regulatory Activities
MIMIC: Medical Information Mart for Intensive Care
NER: named entity recognition
NLP: natural language processing
ORADE: opioid-related adverse drug event
POS: part of speech
PT: preferred term
SPINEL: Supporting Pharmacovigilance by Leveraging Artificial Intelligence Methods to Analyze Electronic Health Records Data
SVM: support vector machine
UMLS: Unified Medical Language System

Edited by K El Emam, B Malin; submitted 12.12.22; peer-reviewed by Q Ye, J Shi, D Chrimes; comments to author 06.02.23; revised version received 29.03.23; accepted 02.06.23; published 18.07.23.

Please cite as:

Sorbello A, Haque SA, Hasan R, Jermyn R, Hussein A, Vega A, Zembrzusi K, Ripple A, Ahadpour M
Artificial Intelligence-Enabled Software Prototype to Inform Opioid Pharmacovigilance From Electronic Health Records: Development and Usability Study
JMIR AI 2023;2:e45000
URL: <https://ai.jmir.org/2023/1/e45000>
doi: [10.2196/45000](https://doi.org/10.2196/45000)
PMID: [37771410](https://pubmed.ncbi.nlm.nih.gov/37771410/)

©Alfred Sorbello, Syed Arefinul Haque, Rashedul Hasan, Richard Jermyn, Ahmad Hussein, Alex Vega, Krzysztof Zembrzusi, Anna Ripple, Mitra Ahadpour. Originally published in JMIR AI (<https://ai.jmir.org/>), 18.07.2023. This is an open-access article distributed under the terms of the Creative Commons Attribution License (<https://creativecommons.org/licenses/by/4.0/>), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR AI, is properly cited. The complete bibliographic information, a link to the original publication on <https://www.ai.jmir.org/>, as well as this copyright and license information must be included.

Original Paper

Self-Supervised Electroencephalogram Representation Learning for Automatic Sleep Staging: Model Development and Evaluation Study

Chaoqi Yang¹, MSc; Cao Xiao², PhD; M Brandon Westover³, MD, PhD; Jimeng Sun¹, PhD

¹Computer Science Department, Carle's Illinois College of Medicine, University of Illinois, Urbana Champaign, Urbana, IL, United States

²Relativity Inc, Chicago, IL, United States

³Harvard Medical School, Boston, MA, United States

Corresponding Author:

Jimeng Sun, PhD

Computer Science Department

Carle's Illinois College of Medicine

University of Illinois, Urbana Champaign

201 N Goodwin Ave

Urbana, IL, 61801

United States

Phone: 1 9142698058

Email: jimeng.sun@gmail.com

Abstract

Background: Deep learning models have shown great success in automating tasks in sleep medicine by learning from carefully annotated electroencephalogram (EEG) data. However, effectively using a large amount of raw EEG data remains a challenge.

Objective: In this study, we aim to learn robust vector representations from massive unlabeled EEG signals, such that the learned vectorized features (1) are expressive enough to replace the raw signals in the sleep staging task, and (2) provide better predictive performance than supervised models in scenarios involving fewer labels and noisy samples.

Methods: We propose a self-supervised model, Contrast with the World Representation (ContraWR), for EEG signal representation learning. Unlike previous models that use a set of negative samples, our model uses global statistics (ie, the average representation) from the data set to distinguish signals associated with different sleep stages. The ContraWR model is evaluated on 3 real-world EEG data sets that include both settings: at-home and in-laboratory EEG recording.

Results: ContraWR outperforms 4 recently reported self-supervised learning methods on the sleep staging task across 3 large EEG data sets. ContraWR also supersedes supervised learning when fewer training labels are available (eg, 4% accuracy improvement when less than 2% of data are labeled on the Sleep EDF data set). Moreover, the model provides informative, representative feature structures in 2D projection.

Conclusions: We show that ContraWR is robust to noise and can provide high-quality EEG representations for downstream prediction tasks. The proposed model can be generalized to other unsupervised physiological signal learning tasks. Future directions include exploring task-specific data augmentations and combining self-supervised methods with supervised methods, building upon the initial success of self-supervised learning reported in this study.

(JMIR AI 2023;2:e46769) doi:[10.2196/46769](https://doi.org/10.2196/46769)

KEYWORDS

physiological signals; electroencephalogram; EEG; sleep staging; sleep; predict; wearable devices; wearable; self-supervised learning; digital health; mHealth; mobile health; healthcare; health care; machine learning

Introduction

Deep learning models have shown great success in automating tasks in sleep medicine by learning from high-quality labeled

electroencephalogram (EEG) data [1]. EEG data are collected from patients wearing clinical sensors, which generate real-time multimodal signal data. A common challenge in classifying physiological signals, including EEG signals, is the lack of enough high-quality labels. This paper introduces a novel

self-supervised model that leverages the inherent structure within large, unlabeled, and noisy data sets and produces robust feature representations. These representations can significantly enhance the performance of downstream classification tasks, such as sleep staging, especially in cases where only limited labeled data are available.

Self-supervised learning (specifically, self-supervised contrastive learning) aims at learning a feature encoder that maps input signals into a vector representation using unlabeled data. Self-supervised methods involve two steps: (1) a *pretrain* step to learn the feature encoder without labels and (2) a *supervised* step to evaluate the learned encoder with a small amount of labeled data. During the pretrain step, some recent methods (eg, Momentum Contrast [MoCo] [2] and the simple framework for contrastive learning of visual representations [SimCLR] [3]) use the feature encoder to construct positive and negative pairs from the unlabeled data and then optimize the encoder by pushing positive pairs closer and negative pairs farther away. A positive pair consists of 2 different augmented versions of the same sample (ie, applying 2 data augmentation methods separately to the same sample), while a negative pair is generated from the augmented data of 2 different samples. For example, the augmentation method for EEG data can be denoising or channel flipping. In this practice, existing negative sampling strategies often incur sampling issues [4,5], especially for noisy EEG data, which significantly affects performance [6]. Specifically, in the self-supervised learning setting (without labels), the negative samples are actually random samples, which may be from the same latent class. Using these “negative samples” can potentially undermine model performance.

Technically, this study contributes to the pretrain step, where we address the aforementioned limitations of existing negative sampling strategies (eg, MoCo [2] and SimCLR [3]) by leveraging global data statistics. In contrastive learning, positive pairs provide similarity-related information, while negative pairs provide contrastive information. Both types of information are essential in learning an effective feature encoder. This study proposes a new contrastive learning method, named Contrast with the World Representation (ContraWR). In our ContraWR, we construct positive pairs using data augmentation, similar to existing methods, while we use one global average representation over the data set (called the *world representation*) as the negative sample to provide the contrastive information. Derived from global data statistics, the world representation is robust even in noisy environments, and it follows a new contrastive guidance in the absence of labels: *the representation similarity between positive pairs is stronger than the similarity to the world representation*. Moreover, in this study, we later strengthen our model with an instance-aware world representation for individual samples, where closer samples

have larger weights in calculating the global average. Our experiments show that the instance-aware world representation makes the model more accurate, and this conclusion aligns with the findings from a previous paper [6] that harder negative samples are more effective in learning feature encoding.

We evaluated the proposed ContraWR on the sleep staging task with 3 real-world EEG data sets. Our model achieved results comparable to or better than those of recent popular self-supervised methods including MoCo [2], SimCLR [3], Bootstrap Your Own Latent (BYOL) [7], and simple Siamese (SimSiam) [8]. The results also show that self-supervised contrastive methods, especially our ContraWR method, are much more powerful in low-label scenarios than supervised learning (eg, 4% accuracy improvement on sleep staging with less than 2% training data of the Sleep EDF data set).

Methods

EEG Data Sets

We considered 3 real-world EEG data sets for this study (the first 2 data sets entirely comprise at-home PSG recordings):

1. The data set of the Sleep Heart Health Study (SHHS) [9,10] is a multicenter cohort study from the National Heart Lung & Blood Institute (Bethesda, Maryland), assembled to study sleep-disordered breathing, which comprises 5804 adult patients older than 40 years and 5445 recordings in the first visit. We used first-visit polysomnography (PSG) data in the experiments. Each recording has 14 PSG channels, and the recording frequency is 125.0 Hz. We used the C3/A2 and C4/A1 EEG channels.
2. The Sleep EDF [11] cassette portion is another benchmark data set collected in a 1987-1991 study of age effects on sleep in healthy Caucasians. The data comprise 78 subjects aged 25-101 years who were taking non-sleep-related medications; the data set contains 153 full-night EEG recordings with a recording frequency of 100.0 Hz. We extracted the Fpz-Cz/Pz-Oz EEG channels as the raw inputs to the model.
3. The Massachusetts General Hospital's (MGH's) MGH Sleep data set [1] was collected from MGH's sleep laboratory, which comprises more than 5000 individuals, where 6 EEG channels (ie, F3-M2, F4-M1, C3-M2, C4-M1, O1-M2, and O2-M1) were used for sleep staging, recorded at a 200.0-Hz frequency. After filtering out mismatched signals and missing labels, we finally curated 6478 recordings.

The data set's statistics can be found in [Table 1](#), and the class label distribution is shown in [Table 2](#).

Table 1. Data set statistics.

Name	Location	Channels, n	Recordings, n	Epochs, n	Storage (GB)
Sleep Heart Health Study	At home	2	5445	4,535,949	260
Sleep EDF	At home	2	153	415,089	20
MGH ^a Sleep	In the laboratory	6	6478	4,863,523	1322

^aMGH: Massachusetts General Hospital.

Table 2. Class label distribution of the data sets.

Name	Epochs, n (%)				
	W	N1	N2	N3	R
Sleep Heart Health Study	1,306,742 (28.8)	169,021 (3.7)	1,856,130 (40.9)	571,191 (12.6)	632,865 (14.0)
Sleep EDF	285,561 (68.8)	21,522 (5.2)	69,132 (16.6)	13,039 (3.2)	25,835 (6.2)
MGH ^a Sleep	2,154,540 (44.3)	481,488 (9.9)	700,347 (14.4)	855,980 (17.6)	671,168 (13.8)

^aMGH: Massachusetts General Hospital.

Problem Formulation

To set up the experiments, the raw subject EEG recordings, which are multichannel brain waves, were used. First, the unlabeled subject recordings were grouped as the *pretrain* set, and the labeled recordings were grouped into the *training* or *test* sets. The training and test sets are usually small, but their EEG recordings are labeled, while the pretrain set contains a large number of unlabeled recordings. Within each set, the long recordings are segmented into disjoint 30-second windows. Each window is called *an epoch*, denoted as $x \in \mathbb{R}^{C \times N}$. Each epoch has the same format: C input channels and N time stamps from each channel.


For these data sets, the ground truth labels were released by the original data publishers. To align with the problem's setting, participants were randomly assigned to the pretrain set, training set, and test set in different proportions (90%: 5%: 5% for the Sleep EDF and MGH sets and 98%: 1%: 1% for the SHHS set, since they have different amounts of data). All epochs segmented from a participant are placed within the same set. The pretrain set is used for self-supervised learning; hence, we removed their labels.





In the pretrain step, the EEG self-supervised representation learning problem requires building a feature encoder $f(\cdot)$ from the pretrain set (without labels), which maps an epoch x into a vector representation $h \in \mathbb{R}^d$, where d is the feature dimensionality, such that the representation h can replace raw signals for downstream classification tasks. Evaluation of the encoder $f(\cdot)$ was conducted on the training and test data (with labels). We focus on sleep staging as the *supervised* step, where the feature vector of a sample x will be mapped to 5 sleep cycle labels, awake (W), rapid eye movement (REM; R), non-REM 1 (N1), non-REM 2 (N2), and non-REM 3 (N3), based on the American Academy of Sleep Medicine's (AASM's) scoring standards [12]. Specifically, based on the feature encoder from the pretrain step, the training set is used to learn a linear model

on top of the feature vectors, and the test set is used to evaluate the linear classification performance.

Background and Existing Methods

Overview

Self-supervised learning occurs in the pretrain step, and it uses representation similarity to exploit the unlabeled signals, with an encoder network $f(\cdot): \mathbb{R}^{C \times N} \rightarrow \mathbb{R}^d$ and a nonlinear projection network $g(\cdot): \mathbb{R}^d \rightarrow \mathbb{R}^m$. Specifically, for a given signal x from the pretrain set, commonly, one applies data augmentation methods $a(\cdot)$ to produce 2 different modified signals $x \sim'$, $x \sim''$ (after this procedure, the format does not change), which are then transformed into h' , $h'' \in \mathbb{R}^d$ by $f(\cdot)$ and further into z' , $z'' \in \mathbb{R}^m$ by $g(\cdot)$. The vectors z' , z'' are finally normalized with the L_2 norm onto the unit hypersphere .

We call  the *anchor*,  the *positive sample*, and these 2 together are called a *positive pair*. For the projections z_k obtained from other randomly selected signals (by negative sampling strategy), their representation  is commonly conceived of as negative samples (though they are random samples), and any one of them together with the anchor is called a *negative pair* in the existing literature [2,3]. The loss function L is derived from the similarity comparison between positive and negative pairs (eg, encouraging the similarity of positive pairs to be stronger than that of all the negative pairs, referred to as the noise contrastive estimation loss [13]). A common forward flow of self-supervised learning on EEG signals can be illustrated as .

For data augmentation, this study used bandpass filtering, noising, channel flipping, and shifting (see the definition in [Multimedia Appendix 1](#) and the visual illustrations in [Multimedia Appendix 2](#)). We conducted ablation studies on the augmentation methods in our experiment and have provided the implementation details. To reduce clutter, we also used z to denote the L_2 normalized version in the rest of the paper.

ContraWR

Background

As mentioned above, most existing models use random samples as negative samples, which can introduce issues (that the negative sample might be from the same latent class) for the pretrain step and undermine representation quality. To address the issue, this paper proposes a new self-supervised learning method, ContraWR. ContraWR replaces the large number of negative samples with a single average representation of the batch, called the world representation or global representation. This way is robust as it avoids constructing negative pairs where 2 data are actually obtained from the same latent class. The world representation serves as a reference in our new contrastive principle: the representation similarity between a positive pair should be stronger than the similarity between the anchor and the world representation. Note that the world representation is not fixed but changes with the encoder updating the parameters.

The World Representation

Assume z' is the anchor, z'' is the positive sample, and z_k denotes a random sample. We generate an average representation of the data set, z_w , as the only negative sample. To formalize, we assume $k \sim p(\cdot)$ is the sample distribution over the data set (ie, k is the sample index), independent of the anchor z' . The world representation z_w is defined by $z_w = E_{k \sim p(\cdot)}[z_k]$.

Here, we denote $D = \{z: \|z\| \leq 1, z \in \mathbb{R}^m\}$. Obviously, $z_w \in D$. In the experiment, z_w is approximated by the average over each batch; that is, we used the average sample representation over the batch \boxed{x} as the world representation, where M is the batch size.

Gaussian Kernel Measure

We adopted a Gaussian kernel defined on D , $sim(x,y): D \times D \rightarrow (0,1]$ as a similarity measure. Formally, given

2 feature projections z', z'' the similarity is defined as \boxed{x} , where σ is a hyperparameter. The Gaussian kernel combined with the following triplet loss gives the alignment and uniformity properties in the loss convergence (Multimedia Appendix 3). When σ becomes large, the Gaussian kernel measure will reduce to cosine similarity.

Loss Function

For the anchor z' , the positive sample z'' and the world representation z_w , we devise a triplet loss, $L = [sim(z', z'') + \delta - sim(z', z_w)]_+$, where $\delta > 0$ is the empirical margin, a hyperparameter. The loss is minimized over batches, ensuring that the similarity of positive pairs $sim(z', z'')$, is larger than the similarity to the world representation $sim(z', z_w)$, by a margin of δ .

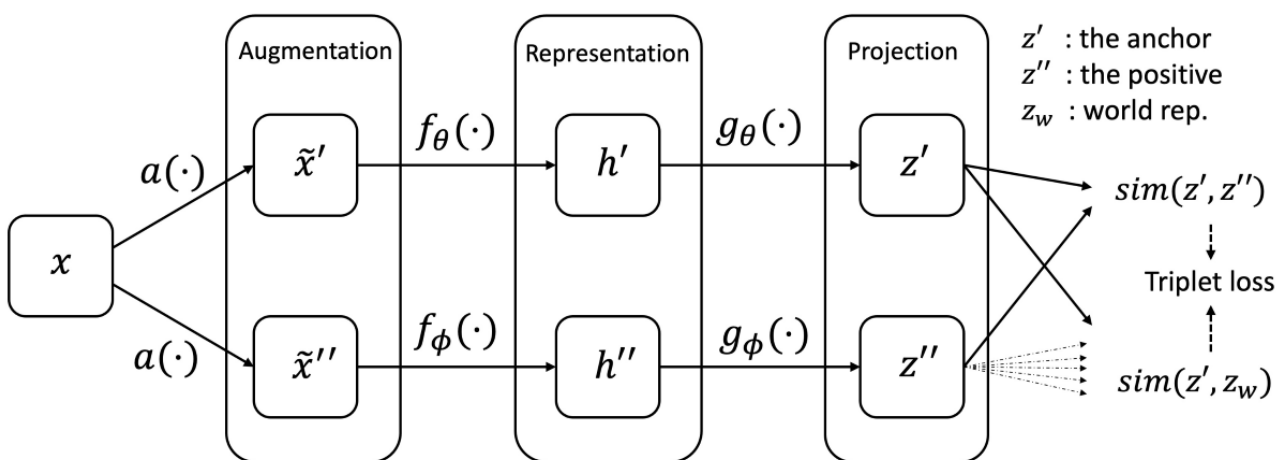
The pipeline of our ContraWR is shown in Figure 1. The online networks $f_\theta(\cdot), g_\theta(\cdot)$ and the target networks $f_\phi(\cdot), g_\phi(\cdot)$ share an identical network structure. Encoder networks $f_\theta(\cdot), f_\phi(\cdot)$ map 2 augmented versions of the same signal to respective feature representations. Then, the projection networks $g_\theta(\cdot), g_\phi(\cdot)$ project the feature representations onto a unit hypersphere, where the loss is defined. During optimization, the web-based networks are updated by gradient descent, and the target networks update parameters from the online network with an exponential moving average (EMA) trick [2].

$$\theta^{(n+1)} \leftarrow \theta^{(n)} - \eta \cdot \nabla_{\theta} L$$

$$\phi^{(n+1)} \leftarrow \lambda \cdot \phi^{(n)} + (1 - \lambda) \cdot \theta^{(n+1)}$$

where n indicates the n th update, η is the learning rate, and λ is a weight hyperparameter. After this optimization in the pretrain step, the encoder network $f_\theta(\cdot)$ is ready to be evaluated on the training and test sets in the supervised step.

Figure 1. The Contrast with the World Representation (ContraWR) model pipeline. We show the 2-way model pipeline in this figure. The web-based network (upper) is updated by gradient descent, while the target network (lower) is updated by the exponential moving average. Finally, the results of the 2 models form the triplet loss function.

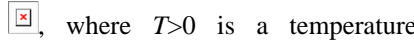
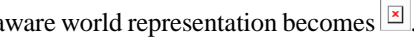


ContraWR+: Contrast With Instance-Aware World Representation

Background

To learn a better representation, we introduced a weighted averaged world representation based on the harder principle: the similarity between a positive pair should be stronger than the similarity between the anchor and the weighted average feature representations of the data set, where the weight is set higher for closer samples. We call the new model ContraWR+. This is a more difficult objective than the simple global average in ContraWR.

Instance-Aware World Representation

In this new model, the world representation is enhanced by modifying the sampling distribution to be instance-specific. We define $p(\cdot|z)$ as the instance-aware sampling distribution of an anchor z , which is different from the sample distribution $p(\cdot)$ used in ContraWR, , where $T > 0$ is a temperature hyperparameter, such that similar samples are selected with higher probability parametrized by $p(\cdot|z)$. Consequently, for an anchor z' , the instance-aware world representation becomes .

Here, T controls the contrastive hardness of the world representation. When $T \rightarrow \infty$, $p(\cdot|z)$ is asymptotically identical to $p(\cdot)$, and the above equation reduces to the simple global average form $z_w = E_{k~p(\cdot)} [z_k]$; while $T \rightarrow 0^+$, the form becomes trivial, $z_w = \text{argmax}_{z_k} (\text{sim}(z', z_k))$. We have tested different T and found that the model is not sensitive to T over a wide range. Here, z_w is also practically implemented by using the weighted average over each batch. We can rewrite the similarity measure given the anchor z_i and the new world representation z_w as:

$$\text{sim}(z_i, z_w) = \text{sim}(z_i, E_{k~p(\cdot|z')} [z_k])$$



In this new method, we also used triplet loss as the final objective.

Implementations

Signal Augmentation

For the experiments, we used four augmentation methods, illustrated in [Multimedia Appendix 2](#): (1) bandpass filtering: to reduce noise, we used an order-1 Butterworth filter (the bandpass is specified in [Multimedia Appendix 2](#)); (2) noising: we added extra high- or low-frequency noise to each channel, mimicking the physical distortion; (3) channel flipping: corresponding sensors from the left side and the right of the head were swapped due to symmetricity; and (4) shifting: within

one sample, we advanced or delayed the signal for a certain time span. Detailed configurations of augmentation methods vary for the 3 data sets, and we have listed them in [Multimedia Appendix 2](#).

Baseline Methods

In the experiments, several recent self-supervised learning methods were implemented for comparison.

MoCo [2] devises 2 parallel encoders with an EMA. It also uses a large memory table to store new negative samples, which are frequently updated.

SimCLR [3] uses an encoder network to generate both anchor and positive samples, where negative samples are collected from the same batch.

BYOL [7] also uses 2 encoders: a web-based network and a target network. They put one more predictive layer on top of the web-based network to predict (reconstruct) the result from the target network, while no negative samples are presented.

SimSiam [8] uses the same encoder networks on 2 sides and also does not use the negative samples.

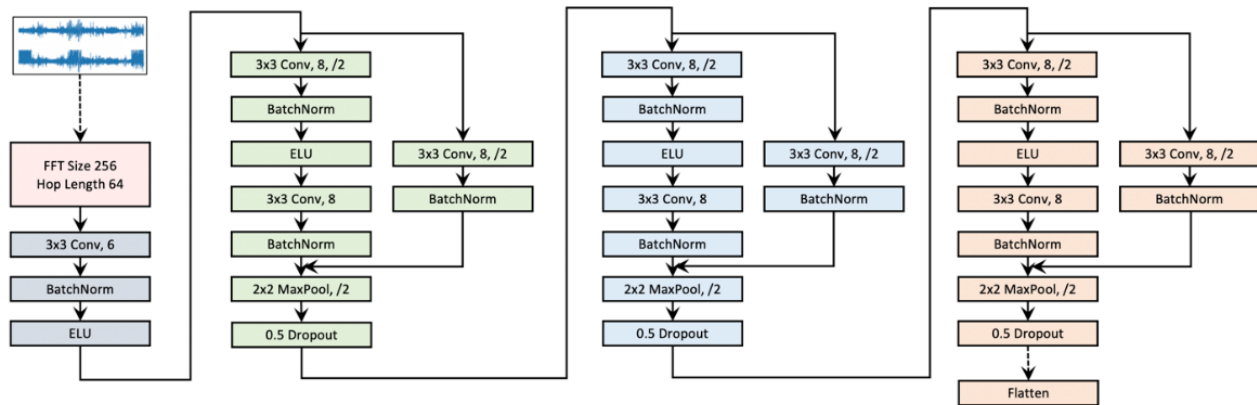
Average k-nearest neighbor TopX is our developed baseline model, which identifies the top X nearest neighbors for each sample within the batch and uses the average representation of these top X neighbors as the negative sample. We used the same triplet loss as our ContraWR model. In the experiments, we tested X=1, X=5, and X=50. When X approaches the batch size, this model will gradually reduce to ContraWR.

Model Architecture

For a fair comparison, all models, including baseline approaches and our models, use the same augmentation and encoder architecture, as shown in [Figure 2](#). This architecture cascades a short-time Fourier transform (STFT) operation, a 2D convolutional neural network layer, and three 2D convolutional blocks. Empirically, we found that the application of neural networks generates better accuracy on the STFT spectrogram of the signals than on the raw signals. The same practices were reported by Yang et al [14,15].

We also considered a supervised model (called *Supervised*) as a reference model, which uses the same encoder architecture and adds a 2-layer fully connected network (128, 256, and 192 units for the Sleep EDF, SHHS, and MGH data sets, respectively) for the sleep staging classification task. The supervised model does not use the pretrain set but is trained from scratch on raw EEG signals in the training set and tested on the test set. We also included an untrained encoder model as a baseline, where the encoder was initialized but not optimized in the pretrain step.

Figure 2. The short-time Fourier transform (STFT) convolutional encoder network. The encoder network first transforms raw signals into spectrogram via STFT, and then a convolutional neural network–based encoder is built on top of the spectrogram. ELU: exponential linear unit; FFT: Fast Fourier Transform; Conv.:convolution operation.



Evaluation Protocol

We evaluated performance on the sleep staging task with overall 5-class classification accuracy. Each experiment was conducted with 5 different random seeds. For self-supervised methods, we optimized the encoder for 100 epochs (here, “epoch” is a concept in deep learning) with unlabeled data, used the training set to find a good logistic classifier, and used the test set data for evaluation in accordance with He et al [2] and Chen et al [3]. For the supervised method, we trained the model for 100 epochs on the training set. Our setting ensures the convergence of all models.

Results

Better Accuracy in Sleep Staging

Comparisons on the downstream sleep staging task are shown in Table 3.

All self-supervised methods outperformed the untrained encoder model, indicating that the pretrain step does learn some useful features from unlabeled data. We observed that ContraWR and ContraWR+ both outperform the supervised model, suggesting that the feature representations provided by the encoder can better preserve the predictive features and filter out noises than using the raw signals for the sleep staging task, in cases when the amount of labeled data available are not sufficient (eg, less than 2% in Sleep EDF). Compared to other self-supervised methods, our proposed model ContraWR+ also provided better predictive accuracy; that is, about 1.3% on Sleep EDF, 0.8% on SHHS, 1.3% on MGH Sleep. The performance improvements were mostly significant ($P < .001$; comparing MoCo vs Sleep EDF data sets, $P = .002$). MGH Sleep data contain more noise than the other 2 data sets (reflected by the relatively low accuracy with the supervised model on raw signals). Performance gain was notably much more significant on MGH over other self-supervised or supervised models (about 3.3% relative improvement on accuracy), which suggests that the proposed models handle noisy environments better.

Table 3. Comparison of sleep staging accuracy with different methods.

Name	Sleep staging accuracy (%), mean (SD) ^a		
	Sleep EDF data set	Sleep Heart Health Study data set	MGH ^b Sleep data set
Supervised	84.98 (0.3562)	75.61 (0.9347)	69.73 (0.4324)
Untrained Encoder	77.83 (0.0232)	60.03 (0.0448)	55.64 (0.0082)
MoCo ^c	85.58 (0.7707)	77.10 (0.2743)	62.14 (0.7099)
SimCLR ^d	83.79 (0.3532)	76.61 (0.3007)	67.32 (0.7749)
BYOL ^e	85.61 (0.7080)	76.64 (0.3783)	70.75 (0.1461)
SimSiam ^f	84.78 (0.8028)	74.25 (0.4796)	62.08 (0.4902)
AVG-KNN-Top1 ^g	80.39 (1.3721)	69.70 (0.8944)	60.73 (0.7423)
AVG-KNN-Top5	83.24 (0.6182)	75.18 (0.7845)	69.14 (0.3393)
AVG-KNN-Top50	86.35 (0.3246)	77.63 (0.3625)	71.95 (0.3482)
ContraWR ^h	85.94 (0.2326)	77.52 (0.5748)	71.97 (0.1774)
ContraWR+	86.90 (0.2288)	77.97 (0.2693)	72.03 (0.1823)

^aCalculated over 5 random seeds.

^bMGH: Massachusetts General Hospital.

^cMoCo: Momentum Control.

^dSimCLR: simple framework for contrastive learning of visual representations.

^eBYOL: Bootstrap Your Own Latent.

^fSimSiam: simple Siamese.

^gAVG-KNN-TopX: average k-nearest neighbor TopX.

^hContraWR: Contrast with the World Representation.

Ablation Study on Data Augmentations

We also inspected the effectiveness of different augmentation methods on EEG signals, shown in [Table 4](#).

We empirically test all possible combinations of 4 considered augmentations: channel flipping, bandpass filtering, noising,

and shifting. Since channel flipping cannot be applied by itself, we combined it with other augmentations. The evaluation was conducted on Sleep EDF data with the ContraWR+ model. To sum up, all augmentation methods are beneficial, and collectively, they can further boost the classification performance.

Table 4. Evaluation accuracy of different augmentations.

Augmentations	Accuracy (%), mean (SD) ^a
Bandpass	84.23 (0.2431)
Noising	83.60 (0.1182)
Shifting	84.65 (0.2844)
Bandpass + flipping	85.77 (0.2337)
Noising + flipping	84.45 (0.1420)
Shifting + flipping	85.13 (0.0558)
Bandpass + noising	85.37 (0.1214)
Noising + shifting	84.78 (0.1932)
Shifting + bandpass	85.25 (0.1479)
Bandpass + noising + flipping	85.76 (0.1794)
Noising + shifting + flipping	85.17 (0.2301)
Shifting + bandpass + flipping	86.38 (0.2789)

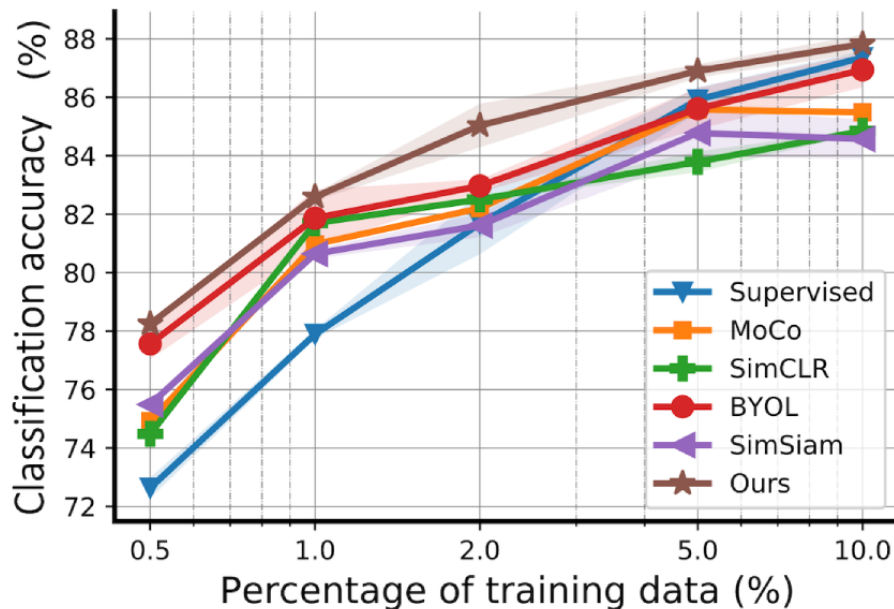
^aCalculated over 5 random seeds.

Varying Amount of Training Data

To further investigate the benefits of self-supervised learning, we evaluated the effectiveness of the learned feature representations with varying training data on Sleep EDF (Figure 3). The default setting is to split all the data into pretrain, training, or test sets by 90%: 5%: 5%. In this section, we maintained the 5% test set constant and resplit the pretrain and training sets (after resplitting, we ensured that all the training set data have labels and removed the labels from the pretrain set), such that the training proportion becomes 0.5%, 1%, 2%, 5%, and 10%, and the rest is used for the pretrain set. This resplitting was conducted at the subject level, after which we

again segmented each subject's recording within the pretrain or training set. We compared our ContraWR+ model to MoCo, SimCLR, BYOL, SimSiam, and the supervised baseline models. Similar ablation studies on SHHS and MGH can be found in Multimedia Appendix 4. Our model outperforms the compared models consistently with different amounts of training data. For example, our model achieves similar performance (with only 5% data as training) to that of the best baseline, BYOL, which needs twice the amount of training data (10% data as training). Also, compared to the supervised model, the self-supervised methods performed better when the labels were insufficient; for example, only $\leq 2\%$ of the data were labeled.

Figure 3. Model performance with different amounts of training data (on the Sleep EDF data set). The curves indicate mean values and shaded areas show the SD of the training/test over 5 random seeds. All models have the same encoder network architecture. For the self-supervised method, we trained a logistic regression model on top of the frozen encoder with the training set, and for the supervised model, we trained the encoder along with the final nonlinear classification layer from scratch with the training set. The proportion of training data is 0.5%, 1%, 2%, 5%, and 10%. Each configuration runs with 5 different random seeds and the error bars indicate the SD over 5 seeds. BYOL: Bootstrap Your Own Latent; MoCo: Momentum Contrast; SimCLR: simple framework for contrastive learning of visual representations; SimSiam: simple Siamese.



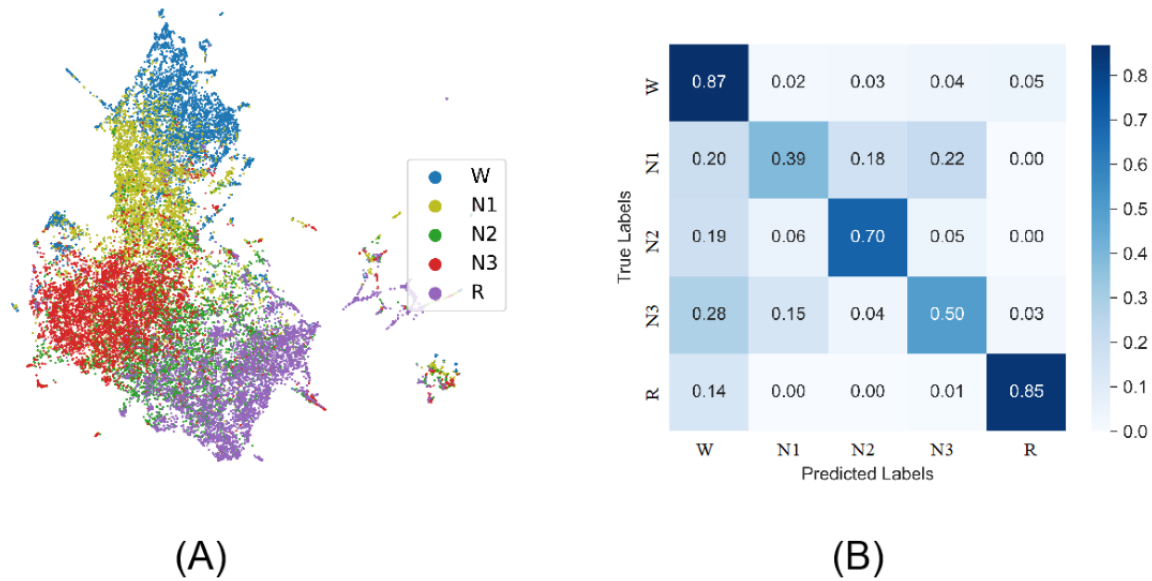
Representation Projection

We next sought to assess the quality of the learned feature representations. To do this, we used the representations produced by ContraWR+ on the MGH data set and randomly selected 5000 signal epochs per label from the data set. The ContraWR+ encoder is optimized on the pretrain step without using the labels. We extracted feature representations for each sample through the encoder network and used uniform manifold approximation and projection (UMAP) [16] to project onto the

2D space. We finally color-coded samples according to sleep stage labels for illustration.

The 2D projection is shown in Figure 4. We also computed the confusion matrix from the evaluation stage (based on the test set; also shown in Figure 4). In the UMAP projection, epochs from the same latent class are closely colocated, which implies that the pretrain step extracts important information for sleep stage classification from the raw unlabeled EEG signals. Stage N1 overlaps with stages W, N2, and N3, which is as expected given that N1 is often ambiguous and thus difficult to classify even for well-trained experts [1].

Figure 4. Uniform manifold approximation and projection and confusion matrix. (A) Using the Massachusetts General Hospital's (MGH's) MGH Sleep data set, we projected the output representations of each signal into a 2D space and color by the actual labels. (B) We have included a confusion matrix on sleep staging.



Hyperparameter Ablation Study

To investigate the sensitivity of our model to hyperparameter settings, we tested with different batch sizes and trained on different values for the Gaussian parameter σ , temperature T , and margin δ . We focused on the ContraWR+ model and evaluated it on the Sleep EDF data set. During the experiment, the default settings are a batch size of 256, σ of 2, T of 2, δ of 0.2, learning rate η of 2×10^{-4} , weight decay of 10^{-4} , and epoch of 100. When testing on 1 hyperparameter, others are maintained constant.

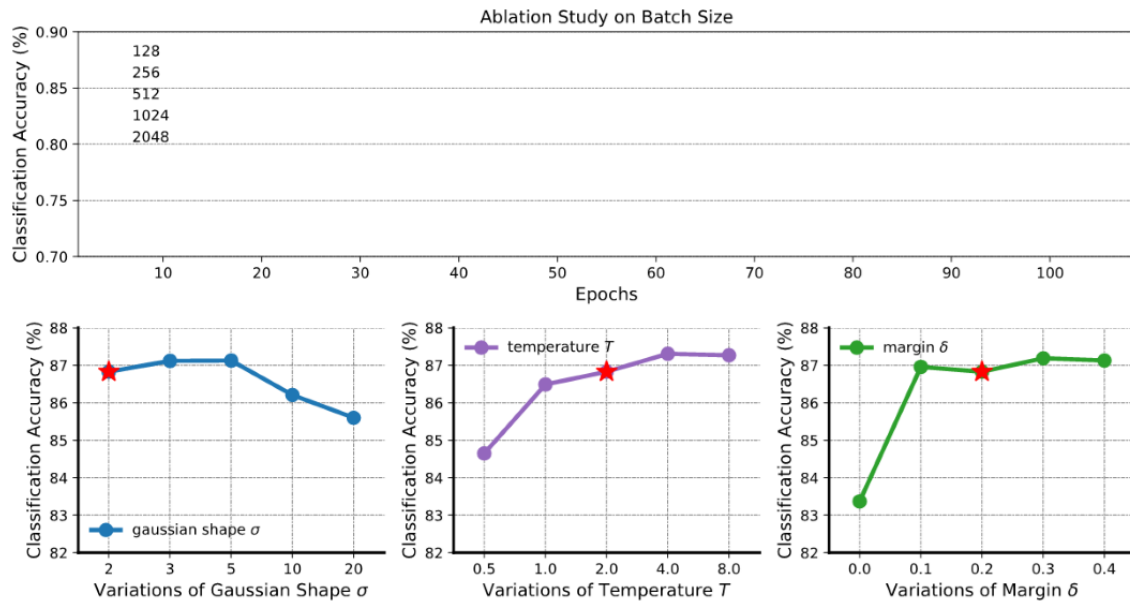
The ablation study's results are shown in Figure 5; the red star indicates the default configuration. Each configuration runs

with 5 different random seeds, and the error bars indicate the SD over 5 experiments. We see that the model is not sensitive to batch size. We see that over a large range (<10) the model is insensitive to the Gaussian width σ . For temperature T , we noted previously that a very small T may be problematic, and a very large T reduces ContraWR+ to ContraWR. Based on the ablation experiments, the performance is relatively insensitive to choices of T . For the margin δ , the difference in distance is bounded (given a fixed σ of 2):



Thus, δ should be large enough; that is, $\delta \geq 0.1$.

Figure 5. Ablation study on batch size and 3 hyperparameters. The curves indicate the mean values and shaded areas show the SD of training/test over 5 random seeds. The red star denotes the default setting. It is obvious that with a larger batch size, the model will perform better, but it is not sensitive to all hyperparameters.



Ethical Considerations

This study has been approved by the Institutional Review Board of Beth Israel Deaconess Medical Center (BIDMC IRB protocol #2022P000417 [Brain Informatics Database]).

Discussion

Principal Results

Our proposed ContraWR and ContraWR+ models outperformed 4 recent self-supervised learning methods on the sleep staging task across 3 large EEG data sets ($P < .001$ in almost all cases). ContraWR+ also superseded supervised learning when fewer training labels were available (eg, a 4% improvement in accuracy when less than 2% of data were labeled). Moreover, the models provided well-separated representative structures in 2D projection.

Comparison With Prior Work

Self-Supervised Learning

Many deep generative methods have been proposed for unsupervised representation learning. They mostly rely on autoencoding [17-19] or adversarial training [20-22]. Mutual information maximization is also popular for compressing input data into a latent representation [23-25].

Recently, self-supervised contrastive learning [2,3,7,8,14] has become popular, where loss functions are devised from representation similarity and negative sampling. However, one recent study [4] highlighted inherent limitations of negative sampling and showed that this strategy could hurt the learned representation significantly [5]. To address these limitations, Chuang et al [5] used the law of total probability and approximated the per-class negative sample distribution using the weighted sum of the global data distribution and the expected

class label distribution. However, without the actual labels, the true class label distribution is unknown. Grill et al [7] and Chen and He [8] proposed ignoring negative samples and learning latent representations using only positive pairs.

In this paper, we leverage the negative information by replacing negative samples with the average representation of the batch samples (ie, the world representation). We argue and provide experiments showing that contrasting with the world representation is more powerful and robust in the noisy EEG setting.

EEG Sleep Staging

Before the emergence of deep learning, several traditional machine learning approaches [26-28] significantly advanced the field using hand-crafted features, as highlighted by Biswal et al [29]. Recently, deep learning models have been applied to various large sleep databases. SLEEPNET [29] built a comprehensive system combining many machine learning models to learn sleep signal representations. Biswal et al [1] designed a multilayer recurrent and convolutional neural network model to process multichannel signals from EEG. To provide interpretable stage prototypes, Al-Hussaini et al [30] developed a SLEEPER model that uses a particular deep learning approach called prototype learning guided by a decision tree to provide more interpretable results. These studies rely on a large set of labeled training data. However, the annotations are expensive, and oftentimes the labeled set is small. In this study, we exploited the large set of unlabeled data to improve the classification, which is more challenging.

Self-Supervised Learning on Physiological Signals

While image [31,32], video [33], language [34,35], and speech [36] representations have benefited from contrastive learning, research on learning physiological signals has been limited [37,38]. Lemkhenter et al [39] proposed phase and amplitude

coupling for physiological data augmentation. Banville et al [40] conducted representation learning on EEG signals, and they targeted monitoring and pathology screening tasks, without using frequency information. Cheng et al [41] learned subject-aware representations for electrocardiography data and tested various augmentation methods. While most of these methods are based on pairwise similarity comparison, our model provides contrastive information from global data statistics, providing more robust representations. Also, we extracted signal information from the spectral domain.

Strengths and Limitations

The strengths of our study are (1) we used 3 real-world data sets collected from different institutes and across different year ranges, and 2 are publicly available; (2) our PSG recordings are diverse and generalizable, including 2 data sets collected at home and 1 collected in the laboratory setting, all having relatively large sizes; (3) we have open-sourced our data processing pipelines and all programs used for this study [42], including the baseline model implementations; and (4) we proposed new data augmentation methods for PSG signals and have systematically evaluated their effectiveness. However, the following limitations of our study should be noted: (1) we fixed

the neural network encoder architecture in the study, which we plan to explore using other models including recurrent neural networks in the future; (2) we have used STFT to extract spectrograms, but we may consider alternative techniques such as wavelet transformation in future; and (3) our current data augmentation methods are based on clinical knowledge, and we aim to investigate data-driven approaches to design more effective methods in the future.

Conclusions

This study is motivated by the need to learn effective EEG representations from large unlabeled noisy EEG data sets. We propose a self-supervised contrastive method, ContraWR, and its enhanced variant, ContraWR+. Instead of creating a large number of negative samples, our method contrasts samples with an average representation of many samples. The model is evaluated on a downstream sleep staging task with 3 real-world EEG data sets. Extensive experiments show that the model is more powerful and robust than multiple baselines including MoCo, SimCLR, BYOL, and SimSiam. ContraWR+ also outperforms the supervised counterpart in label-insufficient scenarios.

Acknowledgments

This work was in part supported by the National Science Foundation (awards SCH-2014438, IIS-1418511, CCF-1533768, and IIS-1838042), the National Institute of Health (R01NS107291, R56HL138415, 1R01NS102190, 1R01NS102574, and RF1AG064312), the Glenn Foundation for Medical Research and the American Federation for Aging Research (Breakthroughs in Gerontology Grant), and the American Academy of Sleep Medicine (AASM Foundation Strategic Research Award).

Authors' Contributions

CY implemented the methods and conducted the experiments. All authors were involved in conceptualizing the study and drafting the manuscript.

Conflicts of Interest

MBW is the cofounder of Beacon Biosignals, which played no role in this study.

Multimedia Appendix 1

Supplementary on model implementation.

[PDF File (Adobe PDF File), 168 KB - ai_v2i1e46769_app1.pdf]

Multimedia Appendix 2

Illustration for data augmentations (bandpass filtering, noising, flipping, and shifting).

[PNG File , 368 KB - ai_v2i1e46769_app2.png]

Multimedia Appendix 3

Theoretical loss boundness analysis.

[PDF File (Adobe PDF File), 562 KB - ai_v2i1e46769_app3.pdf]

Multimedia Appendix 4

Results on SHHS and MGH data set during varying the label sizes.

[PDF File (Adobe PDF File), 24 KB - ai_v2i1e46769_app4.pdf]

References

1. Biswal S, Sun H, Goparaju B, Westover M, Sun J, Bianchi M. Expert-level sleep scoring with deep neural networks. *J Am Med Inform Assoc* 2018 Dec 01;25(12):1643-1650 [FREE Full text] [doi: [10.1093/jamia/ocy131](https://doi.org/10.1093/jamia/ocy131)] [Medline: [30445569](https://pubmed.ncbi.nlm.nih.gov/30445569/)]

2. He K, Fan H, Wu Y, Xie S, Girshick R. Momentum Contrast for Unsupervised Visual Representation Learning. 2020 Presented at: 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR); June 13-19, 2020; Seattle, WA. [doi: [10.1109/cvpr42600.2020.00975](https://doi.org/10.1109/cvpr42600.2020.00975)]
3. Chen T, Kornblith S, Norouzi M, Hinton G. A simple framework for contrastive learning of visual representations. arXiv Preprint posted online February 13, 2020.
4. Arora S, Khandeparkar H, Khodak M, Plevrakis O, Saunshi N. A Theoretical Analysis of Contrastive Unsupervised Representation Learning. arXiv Preprint posted online February 25, 2019. [doi: [10.1090/mbk/121/79](https://doi.org/10.1090/mbk/121/79)]
5. Chuang C, Robinson J, Lin Y, Torralba A, Jegelka S. Debaised Contrastive Learning. 2020 Presented at: 34th Conference on Neural Information Processing Systems (NeurIPS2020); 2020; Vancouver, BC.
6. Robinson J, Chuang C, Sra S, Jegelka S. Contrastive learning with hard negative samples. arXiv Preprint posted online October 9, 2020.
7. Grill J, Strub F, Altché F, Tallec C, Richemond P, Buchatskaya E, et al. Bootstrap Your Own Latent - A New Approach to Self-Supervised Learning. 2020 Presented at: 34th Conference on Neural Information Processing Systems (NeurIPS2020); 2020; Vancouver, BC.
8. Chen X, He K. Exploring Simple Siamese Representation Learning. 2021 Presented at: 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR); June 20-25, 2021; Nashville, TN. [doi: [10.1109/cvpr46437.2021.01549](https://doi.org/10.1109/cvpr46437.2021.01549)]
9. Zhang G, Cui L, Mueller R, Tao S, Kim M, Rueschman M, et al. The National Sleep Research Resource: towards a sleep data commons. *J Am Med Inform Assoc* 2018 Oct 01;25(10):1351-1358 [FREE Full text] [doi: [10.1093/jamia/ocy064](https://doi.org/10.1093/jamia/ocy064)] [Medline: [29860441](https://pubmed.ncbi.nlm.nih.gov/29860441/)]
10. Quan S, Howard B, Iber C, Kiley J, Nieto F, O'Connor G, et al. The Sleep Heart Health Study: design, rationale, and methods. *Sleep* 1997;20(12):1077-1085. [doi: [10.1093/sleep/20.12.1077](https://doi.org/10.1093/sleep/20.12.1077)]
11. Kemp B, Zwinderman A, Tuk B, Kamphuisen H, Oberyé JJ. Analysis of a sleep-dependent neuronal feedback loop: the slow-wave microcontinuity of the EEG. *IEEE Trans Biomed Eng* 2000 Sep;47(9):1185-1194. [doi: [10.1109/10.867928](https://doi.org/10.1109/10.867928)] [Medline: [11008419](https://pubmed.ncbi.nlm.nih.gov/11008419/)]
12. Berry RB, Brooks R, Gamaldo C, Harding SM, Lloyd RM, Quan SF, et al. AASM Scoring Manual Updates for 2017 (Version 2.4). *J Clin Sleep Med* 2017 May 15;13(5):665-666 [FREE Full text] [doi: [10.5664/jcsm.6576](https://doi.org/10.5664/jcsm.6576)] [Medline: [28416048](https://pubmed.ncbi.nlm.nih.gov/28416048/)]
13. Gutmann M, Hyvärinen A. Noise-contrastive estimation: A new estimation principle for unnormalized statistical models. 2010 Presented at: 13th International Conference on Artificial Intelligence and Statistics (AISTATS) 2010; 2010; Sardinia, Italy.
14. Yang C, Qian C, Singh N, Xiao C, Westover M, Solomonik E, et al. ATD: augmenting CP tensor decomposition by self supervision. arXiv Preprint posted online June 15, 2021.
15. Yang C, Westover M, Sun J. ManyDG: many-domain generalization for healthcare applications. arXiv Preprint posted online January 21, 2023.
16. McInnes L, Healy J, Saul N, Großberger L. UMAP: Uniform Manifold Approximation and Projection. *JOSS* 2018 Sep;3(29):861. [doi: [10.21105/joss.00861](https://doi.org/10.21105/joss.00861)]
17. Larochelle H, Bengio Y, Vincent P, Lajoie I, Manzagol PA. Stacked denoising autoencoders: learning useful representations in a deep network with a local denoising criterion. *J Mach Learn Res* 2010;11:3371-3408.
18. Baldi P. Autoencoders, Unsupervised Learning, and Deep Architectures. 2012 Presented at: ICML Workshop on Unsupervised and Transfer Learning; 2012; Bellevue, WA.
19. Kingma D, Welling M. Auto-encoding variational bayes. arXiv Preprint posted online December 20, 2013.
20. Donahue J, Krähenbühl P, Darrell T. Adversarial feature learning. arXiv Preprint posted online May 31, 2016.
21. Goodfellow I, Pouget-Abadie J, Mirza M, Xu B, Warde-Farley D, Ozair S, et al. Generative adversarial networks. *Commun ACM* 2020;63(11):139-144. [doi: [10.1145/3422622](https://doi.org/10.1145/3422622)]
22. Shrivastava A, Pfister T, Tuzel O, Susskind J, Wang W, Webb R. Learning from Simulated and Unsupervised Images through Adversarial Training. 2017 Presented at: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR); July 21-26, 2017; Honolulu, HI. [doi: [10.1109/cvpr.2017.241](https://doi.org/10.1109/cvpr.2017.241)]
23. Hjelm R, Fedorov A, Lavoie-Marchildon S, Grewal K, Bachman P, Trischler A, et al. Learning deep representations by mutual information estimation and maximization. arXiv Preprint posted online August 20, 2018.
24. Tschannen M, Djolonga J, Rubenstein P, Gelly S, Lucic M. On mutual information maximization for representation learning. arXiv Preprint posted online July 31, 2019.
25. Bachman P, Hjelm R, Buchwalter W. Learning representations by maximizing mutual information across views. 2019 Presented at: 33rd Conference on Neural Information Processing Systems (NeurIPS2019); 2019; Vancouver, BC.
26. Fraiwan L, Lweesy K, Khasawneh N, Fraiwan M, Wenz H, Dickhaus H. Classification of sleep stages using multi-wavelet time frequency entropy and LDA. *Methods Inf Med* 2018 Jan 17;49(03):230-237. [doi: [10.3414/me09-01-0054](https://doi.org/10.3414/me09-01-0054)]
27. Anderer P, Moreau A, Woertz M, Ross M, Gruber G, Parapatics S, et al. Computer-assisted sleep classification according to the standard of the American Academy of Sleep Medicine: validation study of the AASM version of the Somnolyzer 24 × 7. *Neuropsychobiology* 2010 Sep 9;62(4):250-264 [FREE Full text] [doi: [10.1159/000320864](https://doi.org/10.1159/000320864)] [Medline: [20829636](https://pubmed.ncbi.nlm.nih.gov/20829636/)]

28. Berthomier C, Drouot X, Herman-Stoica M, Berthomier P, Prado J, Bokar-Thire D, et al. Automatic analysis of single-channel sleep EEG: validation in healthy individuals. *Sleep* 2007 Nov;30(11):1587-1595 [[FREE Full text](#)] [doi: [10.1093/sleep/30.11.1587](https://doi.org/10.1093/sleep/30.11.1587)] [Medline: [18041491](https://pubmed.ncbi.nlm.nih.gov/18041491/)]
29. Biswal S, Kulas J, Sun H, Goparaju B, Westover M, Bianchi M, et al. SLEEPNET: automated sleep staging system via deep learning. arXiv Preprint posted online July 26, 2017.
30. Al-Hussaini I, Xiao C, Westover M, Sun J. SLEEPER: interpretable sleep staging via prototypes from expert rules. 2019 Presented at: 4th Machine Learning for Healthcare Conference; 2019; Ann Arbor, MI.
31. Schroff F, Kalenichenko D, Philbin J. FaceNet: A unified embedding for face recognition and clustering. 2015 Presented at: 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR); June 07-12, 2015; Boston, MA. [doi: [10.1109/cvpr.2015.7298682](https://doi.org/10.1109/cvpr.2015.7298682)]
32. Jing L, Tian Y. Self-supervised visual feature learning with deep neural networks: a survey. *IEEE Trans Pattern Anal Mach Intell* 2021 Nov 1;43(11):4037-4058. [doi: [10.1109/tpami.2020.2992393](https://doi.org/10.1109/tpami.2020.2992393)]
33. Wang J, Jiao J, Liu Y. Self-supervised video representation learning by pace prediction. 2020 Presented at: 16th European Conference on Computer Vision; August 23-28, 2020; Glasgow. [doi: [10.1007/978-3-030-58520-4_30](https://doi.org/10.1007/978-3-030-58520-4_30)]
34. Fang H, Wang S, Zhou M, Ding J, Xie P. CERT: contrastive self-supervised learning for language understanding. arXiv Preprint posted online May 16, 2020. [doi: [10.36227/techrxiv.12308378.v1](https://doi.org/10.36227/techrxiv.12308378.v1)]
35. Mikolov T, Sutskever I, Chen K, Corrado G, Dean J. Distributed representations of words and phrases and their compositionality. arXiv Preprint posted online October 16, 2013.
36. Shukla A, Petridis S, Pantic M. Does visual self-supervision improve learning of speech representations for emotion recognition? *IEEE Trans Affective Comput* 2023 Jan 1;14(1):406-420. [doi: [10.1109/taffc.2021.3062406](https://doi.org/10.1109/taffc.2021.3062406)]
37. Franceschi J, Dieuleveut A, Jaggi M. Unsupervised scalable representation learning for multivariate time series. arXiv Preprint posted online January 30, 2019.
38. Oord A, Li Y, Vinyals O. Representation learning with contrastive predictive coding. arXiv Preprint posted online July 10, 2018.
39. Lemkhenter A, Favaro P. Boosting generalization in bio-signal classification by learning the phase-amplitude coupling. 2020 Presented at: 42nd DAGM GCPR: DAGM German Conference on Pattern Recognition; September 28-October 1, 2020; Tübingen, Germany. [doi: [10.1007/978-3-030-71278-5_6](https://doi.org/10.1007/978-3-030-71278-5_6)]
40. Banville H, Chehab O, Hyvärinen A, Engemann D, Gramfort A. Uncovering the structure of clinical EEG signals with self-supervised learning. *J Neural Eng* 2021 Mar 31;18(4):046020. [doi: [10.1088/1741-2552/abca18](https://doi.org/10.1088/1741-2552/abca18)] [Medline: [33181507](https://pubmed.ncbi.nlm.nih.gov/33181507/)]
41. Cheng JY, Azemi E, Goh H, Dogrusoz KE, Tuzel CO. Subject-aware contrastive learning for biosignals (US Patent US20210374570A1). Google Patents. 2021. URL: [https://patents.google.com/patent/US20210374570A1/en?q=\(Subject-aware+contrastive+learning+biosignals\)&oq=Subject-aware+contrastive+learning+for+biosignals](https://patents.google.com/patent/US20210374570A1/en?q=(Subject-aware+contrastive+learning+biosignals)&oq=Subject-aware+contrastive+learning+for+biosignals) [accessed 2023-06-26]
42. Open EEG Data Preprocessing and SSL Baselines. GitHub. 2023. URL: <https://github.com/ycq091044/ContraWR> [accessed 2023-06-26]

Abbreviations

- AASM:** American Academy of Sleep Medicine
- BYOL:** Bootstrap Your Own Latent
- ContraWR:** Contrast with the World Representation
- EEG:** electroencephalogram
- EMA:** exponential moving average
- MGH:** Massachusetts General Hospital
- MoCo:** Momentum Contrast
- PSG:** polysomnography
- REM:** rapid eye movement
- SHHS:** Sleep Heart Health Study
- SimCLR:** simple framework for contrastive learning of visual representations
- SimSiam:** simple Siamese
- STFT:** short-time Fourier transform
- UMAP:** uniform manifold approximation and projection

Edited by K El Emam, B Malin; submitted 24.02.23; peer-reviewed by J Wen, N Mungoli; comments to author 14.05.23; revised version received 27.05.23; accepted 02.06.23; published 26.07.23.

Please cite as:

Yang C, Xiao C, Westover MB, Sun J

Self-Supervised Electroencephalogram Representation Learning for Automatic Sleep Staging: Model Development and Evaluation Study

JMIR AI 2023;2:e46769

URL: <https://ai.jmir.org/2023/1/e46769>

doi: [10.2196/46769](https://doi.org/10.2196/46769)

PMID: [38090533](https://pubmed.ncbi.nlm.nih.gov/38090533/)

©Chaoqi Yang, Cao Xiao, M Brandon Westover, Jimeng Sun. Originally published in JMIR AI (<https://ai.jmir.org>), 26.07.2023. This is an open-access article distributed under the terms of the Creative Commons Attribution License (<https://creativecommons.org/licenses/by/4.0/>), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR AI, is properly cited. The complete bibliographic information, a link to the original publication on <https://www.ai.jmir.org/>, as well as this copyright and license information must be included.

Original Paper

Enabling Early Health Care Intervention by Detecting Depression in Users of Web-Based Forums using Language Models: Longitudinal Analysis and Evaluation

David Owen¹, MSc; Dimosthenis Antypas¹, MSc; Athanasios Hassoulas², PhD; Antonio F Pardiñas³, PhD; Luis Espinosa-Anke¹, PhD; Jose Camacho Collados¹, PhD

¹School of Computer Science and Informatics, Cardiff University, Cardiff, United Kingdom

²Centre for Medical Education, School of Medicine, Cardiff University, Cardiff, United Kingdom

³Centre for Neuropsychiatric Genetics and Genomics, School of Medicine, Cardiff University, Cardiff, United Kingdom

Corresponding Author:

David Owen, MSc
School of Computer Science and Informatics
Cardiff University
Abacws
Senghennydd Road
Cardiff, CF24 4AG
United Kingdom
Phone: 44 (0)29 2087 4812
Email: owendw1@cardiff.ac.uk

Abstract

Background: Major depressive disorder is a common mental disorder affecting 5% of adults worldwide. Early contact with health care services is critical for achieving accurate diagnosis and improving patient outcomes. Key symptoms of major depressive disorder (depression hereafter) such as cognitive distortions are observed in verbal communication, which can also manifest in the structure of written language. Thus, the automatic analysis of text outputs may provide opportunities for early intervention in settings where written communication is rich and regular, such as social media and web-based forums.

Objective: The objective of this study was 2-fold. We sought to gauge the effectiveness of different machine learning approaches to identify users of the mass web-based forum Reddit, who eventually disclose a diagnosis of depression. We then aimed to determine whether the time between a forum post and a depression diagnosis date was a relevant factor in performing this detection.

Methods: A total of 2 Reddit data sets containing posts belonging to users with and without a history of depression diagnosis were obtained. The intersection of these data sets provided users with an estimated date of depression diagnosis. This derived data set was used as an input for several machine learning classifiers, including transformer-based language models (LMs).

Results: Bidirectional Encoder Representations from Transformers (BERT) and MentalBERT transformer-based LMs proved the most effective in distinguishing forum users with a known depression diagnosis from those without. They each obtained a mean F_1 -score of 0.64 across the experimental setups used for binary classification. The results also suggested that the final 12 to 16 weeks (about 3-4 months) of posts before a depressed user's estimated diagnosis date are the most indicative of their illness, with data before that period not helping the models detect more accurately. Furthermore, in the 4- to 8-week period before the user's estimated diagnosis date, their posts exhibited more negative sentiment than any other 4-week period in their post history.

Conclusions: Transformer-based LMs may be used on data from web-based social media forums to identify users at risk for psychiatric conditions such as depression. Language features picked up by these classifiers might predate depression onset by weeks to months, enabling proactive mental health care interventions to support those at risk for this condition.

(JMIR AI 2023;2:e41205) doi:[10.2196/41205](https://doi.org/10.2196/41205)

KEYWORDS

mental health; depression; internet; natural language processing; transformers; language models; sentiment

Introduction

Background

Major depressive disorder (MDD) is one of the most prevalent mental illnesses worldwide, affecting nearly 5% of adults [1]. Depressive episodes, which are symptoms of MDD and other psychiatric conditions, are even more common, with nearly 30% of individuals developing them at least once in their lifetime [2]. The characteristics of MDD and depressive episodes (“depression” hereafter) include low mood, feelings of worthlessness or guilt, and recurrent thoughts of death [3]. Early intervention has been reported to significantly improve patient outcomes and reduce the financial burden on health care services [4]. However, the stigma associated with psychiatric conditions, such as depression, leads to patients underreporting to health care services [5,6].

Given that a number of individuals who would normally meet the criteria for depression underreport to health care services, consideration should be given to how key symptoms may manifest in written language on social media platforms [7]. Longhand discussion websites such as Reddit are a rich source of such information where users may publish a series of posts spanning many months or years [8]. Natural language processing (NLP) can be used to identify features in posts that are predictive of a user who may have depression. Crucially, if affected users are identified before formal diagnosis, this may provide an opportunity for early health care intervention in these cases.

In this study, we derive a specialized subset of an annotated data set that contains Reddit posts belonging to users who have received a diagnosis of depression. This subset allowed us to consider posts before each user’s approximate diagnosis date.

We used state-of-the-art, domain-specific language models (LMs) to assist in the detection of depression. These LMs outperformed the baseline approaches in various experimental settings. Notably, they are adept at early detection of depression. Moreover, through our model analysis, we provide an exhaustive analysis of the temporal aspect related to preemptive detection, providing insights into the time depression symptoms materialized before the diagnosis. Finally, we investigated the role of sentiment in depressed users’ posts and provided a qualitative analysis based on the model performance.

Related Work

There is a growing body of literature on the use of NLP techniques to analyze depression patterns on social media [9,10].

Yates et al [11] developed an approach to distinguish forum users who self-reported a diagnosis of depression from those who did not. It used a convolutional neural network to aggregate user posts in a purpose-built data set, the Reddit Self-reported Depression Diagnosis (RSDD) data set. Their follow-up work involved the conception of a sister data set, RSDD-Time [12], which contained Reddit posts where users declared a past diagnosis of depression, and this diagnosis was linked to an estimated date. Dates were inferred from explicit but often imprecise time expressions in user posts. However, these works did not consider the preemptive detection of depression among

Reddit users in their data sets. That is, they did not consider methods for detecting depression in users before their diagnoses.

Recent NLP studies have explicitly focused on the early detection of depression. Preemptive detection of mentions of depression among Twitter users has been demonstrated with a degree of success by Owen et al [13]. Abed-Esfahani [14] reported similar findings using Reddit data. However, both studies were limited by the uncertainty of whether the users referring to this condition were formally diagnosed. Shah et al [15] also considered approaches for the early detection of depression in Reddit users. In this case, it was determined whether the user had received a physician’s diagnosis. However, it was not certain whether the users’ posts occurred before or after their diagnoses because the dates of the diagnoses were unknown. To gauge the effectiveness of the preemptive detection methods, a series of user posts before a known diagnosis date is required. Eichstaedt et al [16] examined the language in Facebook posts that may have been predictive of depression, as shown in patients’ medical records. They achieved an F_1 -score of 0.66 via logistic regression modeling, which used only the language preceding each patient’s depression diagnosis.

Therefore, this study also sought to extend existing work on preemptive depression detection. We considered social media users whose depression diagnosis date is known and used LMs to harness the language of user posts.

Ren et al [17] performed emotion-driven detection of depression using Reddit, achieving F_1 -scores exceeding 0.9. Their work considered individual depression posts, rather than a series of posts. Nevertheless, their effective use of emotional semantic information suggested that the dissection of our own results could be enhanced using sentiment analysis, which we included in our analysis to provide further insights.

Objectives

We sought to gauge the performance of several machine learning classifiers in the task of distinguishing between RSDD data set users reporting and not reporting a diagnosis of depression, which from here onward we will term as “depressed” and “controls,” respectively. We then used the best-performing classifier in a temporally driven binary classification task. The purpose was to determine the volume of posts in a depressed user’s post timeline, which was the most indicative of their illness. To do this, we considered only the posts authored before the depressed users’ estimated diagnosis dates. Moreover, we considered only posts published up to 6 months before those dates.

The motivation for considering this 6-month time range hails from Winkour et al [18], and their observation that over 50% of patients with depression experienced their first onset at least 6 months before their formal diagnosis. Reece et al [19] made similar observations when examining Twitter users.

The time during which individuals with symptoms or traits of depression remain undiagnosed poses serious health risks. Patients who remain undiagnosed and thus untreated experience a worse outcome than would be the case if they were treated [20], particularly after their first episode [21]. Methods for

assessing suitable time points for health care interventions are needed to identify ways to improve patient outcomes. They are also likely to advance the field of psychiatric therapeutics by supporting modifications to clinical guidelines or the design of randomized controlled trials [22]. A larger body of evidence on this matter could also help identify patients to be targeted for more thorough mental health assessments and provided with further resources, support, and treatment [23].

Methods

Data Description

Overview

Our work is based on the RSDD and RSDD-Time data sets [24]. The RSDD contains Reddit posts of 9210 depressed users and 108,731 control users. The posts were published between

January 2006 and October 2016. The representation of users in RSDD is presented in [Textbox 1](#).

RSDD-Time contains 598 annotated Reddit posts, each of which belongs to a user who declares that they have been formally diagnosed with depression. The posts were published between June 2009 and October 2016. Of these posts, 529 belonged to depressed users that were also present in the RSDD.

RSDD-Time annotations include the recency of a user's diagnosis with respect to the date on which their post was authored. The permissible recency annotations are as follows:

0, unspecified; 1, in the past; 2, up to 2 months ago; 3, between 2 months and 1 year ago; 4, between 1 and 3 years ago; and 5, more than 3 years ago.

The representation of users in RSDD-Time is depicted in [Textbox 2](#).

Textbox 1. An abstract representation of Reddit Self-reported Depression Diagnosis user data. It is not permissible to reveal true user IDs, post dates, or post texts due to privacy reasons.

```
{user_id: 1, posts: [ (<date 1>, <text>),..., (<date n>, <text> ) ], label: <either depressed or control>},
{user_id: 2, posts: [ (<date 1>, <text>),..., (<date n>, <text> ) ], label: <either depressed or control>},
...,
{user_id: n, posts: [ (<date 1>, <text>),..., (<date n>, <text> ) ], label: <either depressed or control>}
```

Textbox 2. An abstract representation of Reddit Self-reported Depression Diagnosis–Time user data. It is not permissible to reveal true user IDs, diagnosis post texts, or post dates, due to privacy reasons.

```
{user_id: 1, diagnosis_post: <text>, post_date: <date>, recency: <0, 1, 2, 3, 4, or 5>},
{user_id: 2, diagnosis_post: <text>, post_date: <date>, recency: <0, 1, 2, 3, 4, or 5>},
...,
{user_id: n, diagnosis_post: <text>, post_date: <date>, recency: <0, 1, 2, 3, 4, or 5>}
```

Deriving RSDD-Matched

We used this information to estimate the diagnosis dates of the 529 users present in both RSDD and RSDD-Time. Those with recency annotations of 0 or 1 were ignored because their diagnosis dates could not be estimated with any degree of accuracy. For each of the remaining users, we determined whether the estimated diagnosis date fell between the date of their first RSDD post and the date of their RSDD-Time diagnosis post. A total of 72 depressed users remained in the study.

A total of 10 matching control users were sought for each of the 72 depressed users. To accomplish this, candidate control users were randomly retrieved from the RSDD and analyzed sequentially. The candidates' posts dated before the corresponding depressed user's estimated diagnosis date were considered. If the number of posts belonging to the candidate

did not vary by >15% with respect to the depressed user, the candidate was considered a match. A control user matched in this manner was not considered a candidate for subsequent depressed users.

Because sufficient matching control users could not be found for 2 of the depressed users, they were excluded from the resulting data set. The data set contained 70 depressed users, each of whom had 10 matching control users. Thus, there were a total of 770 users. The posts were published between April 2006 and June 2016. We named our data set RSDD-Matched. The characteristics of RSDD-Matched are shown in [Table 1](#). Statistics pertaining to individual users in RSDD-Matched can be found in [Multimedia Appendix 1](#).

Because RSDD does not include posts made in mental health subreddits, a depressed user's diagnosis is certain to not be revealed until the time of their diagnosis post. There is language indicative of mental health conversation in the other subreddits.

Table 1. Statistics of the Reddit Self-reported Depression Diagnosis–Matched data set.

	Depressed users	Control users
Total users	70	700
Total posts	36,826	364,747
Total words	1,742,388	8,188,090
Average posts per user	526.1	521.1
Average words per post	47.3	22.4
Shortest post (words)	1	1
Longest post (words)	2642	1894

Descriptive Analysis of RSDD

To better understand our data set, we performed a simple descriptive analysis of RSDD. Word-level exploratory analyses of corpora have been extensively used in corpus linguistics and NLP to gain insight into word prominence. Typically, these follow a bag-of-words [25], pointwise mutual information [26], or term frequency–inverse document frequency (TF-IDF) [27] approach. In our case, we used lexical specificity [28], which is a statistical measure based on hypergeometric distribution, to identify the most prominent words in a corpus. We chose to use lexical specificity because it is structured in a way that is ideal for extracting corpus-specific vocabulary given a global corpus (RSDD) and its subsets (depressed and control users) [29]. It is also a more robust metric for term importance when dealing with different lengths of text [30], which is often the case for Reddit posts.

RSDD is partitioned into 2 subsets, or subcorpora, one containing posts of depressed users, and another containing posts of the control users. After lemmatizing the corpus, lexical specificity analysis revealed the unigrams (single words) that were the most frequently used by depressed and control participants (Table 2). The score column indicates the relevance of a unigram to each subset. For reference, the term “woman”

makes up 0.18% (460,893/257,873,124) of the total words that appear in the depressed user subset compared with only 0.06% (569,330/950,988,726) of the control user subset.

To put the results into context, we should mention that a lexical specificity score of X for a given word W with frequency f means that the probability of W occurring at least f times in the subcorpus is lower than 10^{-X} (assuming a random distribution). For instance, a lexical specificity score of 42,234 for “game” means that the probability of “game” having a frequency of $f=5,373,938$ or higher in the control users subcorpus is $10^{-42,234}$ (ie, an exceptionally low probability which means “game” is overrepresented in the control users’ subset). In general, we can observe a pattern in which depressed users tend to use more relationship or family-related words (eg, “woman” or “relationship”) and words related to the depression symptoms themselves (eg, “life”). In contrast, control users seem to use more mundane terms related to the subreddit communities, such as game-related terms (eg, “game” or “team”). Although this analysis is based only on the statistical frequency of the terms used, it may provide further evidence that developing automatic methods to identify users with depression may indeed be feasible. In the *Results* section, we extend this initial inspection to better understand the errors made by the automatic models.

Table 2. Top ranked words of Reddit Self-reported Depression Diagnosis depressed and control users in terms of lexical specificity.

User, word	Score
Depressed users	
people	338,131.45
know	164,368.51
thing	150,440.49
feel	118,483.23
time	97,250.09
woman	96,165.35
go	79,611.79
want	75,379.17
life	67,769.01
relationship	62,606.64
Control users	
game	42,234.94
trade	39,445.65
key	30,031.17
team	24,333.73
play	17,389.38
player	16,186.61
shiny	14,032.27
hatch	13,265.87
thank	10,177.49
add	10,005.14

Methodology

In this section, we provide more details of our proposed methods for tackling the depression detection task. Framing the task as a machine learning problem, we considered 9 methods based on linear classifiers and more recent LMs.

The initial baselines entailed a support vector machine (SVM) architecture. SVM is an algorithm that learns by example to assign labels to objects [31]. In our case, the objects are Reddit users, and permissible labels are “depressed” and “control.” SVMs have demonstrated effectiveness in the detection of depression-related posts in Reddit [8,32]. Our SVM configurations used different features derived from user posts. These features included TF-IDF, word embeddings, and a combination of both TF-IDF and word embeddings. The TF-IDF [33] features represent the words deemed most notable among the user posts. Word embedding is a real-valued vector representation of a word [34]. Words with similar meanings have vectors with similar values.

The SVM model used was that of scikit-learn [35], as was the TF-IDF vectorizer implementation. The word embeddings generated for each Reddit post were drawn from global vectors trained on Wikipedia and Gigaword data [36]. These vectors had a dimensionality of 300, similar to the average embedding generated. We performed Reddit posttext preprocessing before

their input to the SVM. All posts underwent quotation normalization; therefore, each quotation character was represented by a single apostrophe. All new lines and carriage return characters were replaced with spaces so that posts were represented as a single line string. The posts were then concatenated on a per-user basis so that each user’s posting history was represented as a single-line string. SVM used a linear kernel, which is appropriate for text-classification problems [37-39].

The remaining 6 classifiers were transformer-based LMs. LMs are a statistical means of predicting words [40], whereas transformers provide a neural-network-based approach to generating such models [41]. Transformer-based LMs have proven effective in detecting psychiatric illness-related Reddit posts [12,42,43]. Therefore, we chose to use transformer-based LMs to support the detection of depression in RSDD-Matched. We chose Bidirectional Encoder Representations from Transformers (BERT) [44] and A Lite BERT (ALBERT) [45], which are appropriate for a wide variety of applications. We also chose 4 specialist LMs: BioBERT [46], Longformer [47], MentalBERT [48], and MentalRoBERTa [48]. BioBERT is suitable for use where biomedical concepts are prevalent, such as electronic medical records [49], patient descriptions [50], and health-related Twitter posts [51]. Longformer is designed for use when text is formed from long documents. Indeed, there

were posts in RSDD-Matched that exceed 2000 words. Finally, MentalBERT and MentalRoBERTa are customized for the domain of mental health care and trained using text drawn from mental health discussion forums.

All 6 transformer-based LMs were pretrained bidirectional language representations. This means that for any given word in a text segment, its neighboring words to both the left and right are examined so that the context of the word is well understood. These representations lend themselves to high performance in text classification tasks when compared with traditional approaches using SVMs, for example [52,53].

We used the Simple Transformers software library [54] to deploy LMs. The library provides an application programming interface to the transformer library, which itself provides access to the BERT, ALBERT, BioBERT, Longformer, MentalBERT, and MentalRoBERTa models [55]. The BERT, ALBERT, BioBERT, Longformer, MentalBERT, and MentalRoBERTa classifiers used were “bert-base-uncased,” “albert-base-v1,” “biobert-base-cased-v1.1,” “longformer-base-4096,” “mental-bert-base-uncased,” and “mental-roberta-base,” respectively. In addition to the default hyperparameters of the Simple Transformers, the LM classifiers were instantiated, with the sliding window enabled. Transformer-based LMs may consume only a limited number of tokens (512 tokens). Because the posting histories of most users in RSDD-Matched exceed 512 words, a specialist approach to applying LMs to these posts is needed. Sliding window is one such approach [56].

Experimental Setup

Preemptive Depression Identification Experiment

The first experiment examined the performance of several machine learning classifiers in the task of distinguishing between

depressed and control users in RSDD-Matched. The purpose of this experiment was to understand the extent to which the preemptive detection of depression in social media is possible. Moreover, this experiment was aimed at understanding the capabilities of machine learning classifiers for this task and the suitability of different methods in the task. The results were used to provide a competitive model for subsequent fine-grained temporal experiments.

We used 9 different classifiers. Three entailed an SVM, as described in the *Methodology* section. The remaining 6 were BERT, ALBERT, BioBERT, Longformer, MentalBERT, and MentalRoBERTa, which are also described in the *Methods* section.

In addition to the aforementioned classifiers, we included a naive baseline that predicted positive instances in all cases.

Because the number of positive instances (ie, depressed users) in RSDD-Matched was small, we chose not to use a traditional train-test split. Instead, we used 5-fold cross-validation; an approach also used by Eichstaedt et al [14]. Furthermore, we varied the number of matching control users across the 4 iterations of the experiment (Table 3).

The purpose of these variations is to test the performance of classifiers against increasingly imbalanced data sets. This mimics the conditions likely to be observed in web-based forums where the number of positive instances (ie, depressed users) is dwarfed by the number of negative instances (ie, nondepressed users).

Table 3. Variations of the preemptive depression identification experiment in terms of the number of matching control users considered.

	Depressed users	Matching control users per depressed user	Total users
Variation 1	70	1	140
Variation 2	70	3	280
Variation 3	70	5	420
Variation 4	70	10	770

Temporal Experiment

The purpose of the second primary experiment was to determine which posting period in a depressed user’s post timeline was the most indicative of depression. This involved the use of a subset of RSDD-Matched users. The performance of binary classifiers versus temporal subsets of the posts in the 6 months before the users’ estimated diagnosis dates was measured.

The RSDD-Matched subset contained only depressed users who had at least one post in the 2 weeks before their estimated diagnosis date. Of the 70 depressed users in our RSDD subset, 14 did not have any posts in this 2-week period. Consequently, we used only 56 depressed users in the temporal experiment. Furthermore, not all 10 control users matched with each of the 56 depressed were useable because some did not have at least one post in this 2-week period. Thus, we performed additional

random exclusions of controls to rebalance the data set. After these exclusions, the data set used in the temporal experiment contained 56 depressed users, each of which had 3 matching control users, totaling to 224 users.

The results of the preemptive depression identification experiment were used to partially inform the design of the temporal experiment. Because BERT scored the highest average F_1 -score across all runs of the preemptive depression identification experiment, it was decided that this was the sole general-purpose transformer-based LM to be used in the temporal experiment. Likewise, MentalBERT had the highest average F_1 -score; therefore, it was selected as the sole specialist LM. The 3 variations of the SVM classifier used in the preemptive depression-identification experiment were used once again.

Once again, we used 5-fold cross-validation. Two chief variations of the RSDD-Matched subset and several different temporal configurations were used (Table 4).

The 2 chief strands to our experimental setup are summarized in Figure 1.

We complemented the temporal experiment with sentiment analysis. The purpose of this study was to identify whether there is a link between sentiment and depression with respect to user posts. Text sentiment has been extensively used as a predictor for detecting signs of depressive mood in microblog users [57-59]. Specifically, negatively charged text has often been correlated with depression via expressions of low mood and suicidal ideation [60]. Approaches used to extract sentiment from social media posts include the use of LMs [61] and lexicons such as Valence Aware Dictionary and Sentiment Reasoner (VADER) [62].

To determine whether there is a relationship between sentiment and depression, we used BERTweet-sentiment, a state-of-the-art transformer model, to classify each post in RSDD-Matched as either negative, neutral, or positive. BERTweet-sentiment is based on the BERTweet [63] implementation, which is trained on a large Twitter corpus and fine-tuned for sentiment analysis. Although the model is not trained on Reddit data, we believe that there are enough overlapping lexical characteristics between the 2 domains in terms of internet slang and text lengths that justify its use.

Our sentiment analysis focused on changes in the sentiment distribution of depressed and control users over time. In step

with the design of our temporal experiment, each user's posts are divided into 6 temporal bands, namely 0-4, 4-8, 8-12, 12-16, 16-20, and 20-24 weeks before their estimated diagnosis date (for a control user, this is the estimated diagnosis of its matched depressed user). The average percentage of each sentiment in each band was considered.

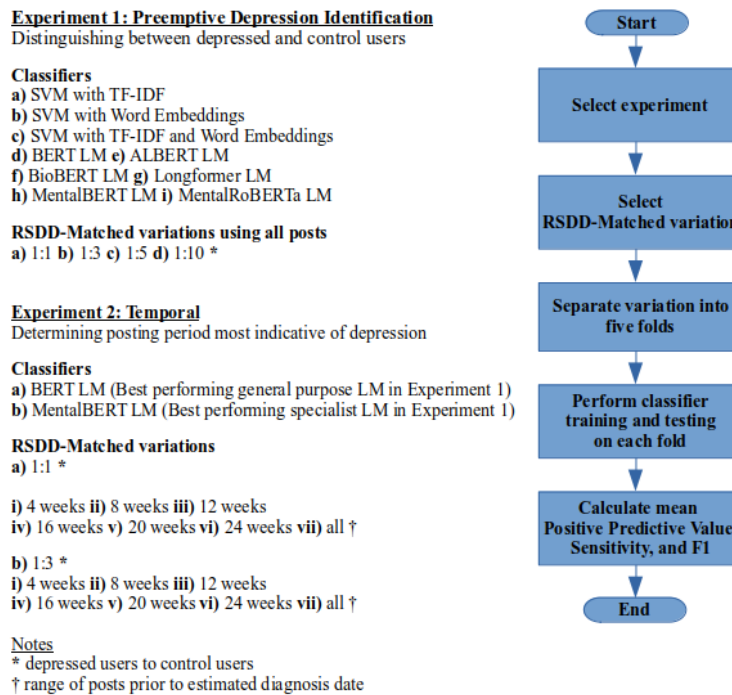
To establish whether the diagnosis was associated with the sentiment of a post, 2 regression models were used. The first was based on the *lme4* framework [64], and the second on *mgcv* [65]. The implementations used were those of the R (version 4.02) statistical environment [66]. We set our outcome variable to be whether a post is "sentimental" (that is, either negative or positive) or not (neutral), and a logistic mixed effects regression was fitted using all the available posts with the individual user identifier as a random effect term. As fixed effects, we used the estimated depression diagnosis (ie, either depressed or control), the time to estimated diagnosis in weeks, the post's word count, and the interaction term of estimated diagnosis with time.

Having sought to establish whether the diagnosis of the user was associated with the sentimentality inferred for each post, we also considered a more fine-grained multinomial regression model. This is equivalent to fitting a series of logistic models against a reference category [67] and is similar to the "stacked" designs used in other disciplines [68]. For our purposes, we will consider "neutral" as the reference category of our multinomial outcome, so all effect sizes will indicate the probability of a post being positive or negative *instead of* neutral.

Table 4. Variations of the temporal experiment in terms of the number of matching control users and numbers of weeks of posts before estimated diagnosis dates considered.

	Depressed users	Matching control users per depressed user	Total users	Weeks of posts included before estimated diagnosis date
Variation 1	56	1	112	4, 8, 12, 16, 20, and 24
Variation 2	56	3	224	4, 8, 12, 16, 20, and 24

Figure 1. Summary of the 2 chief experimental setups. ALBERT: A Lite Bidirectional Encoder Representations from Transformers; BERT: Bidirectional Encoder Representations from Transformers; LM: language model; SVM: support vector machine; TF-IDF: term frequency–inverse document frequency; RSDD: Reddit Self-reported Depression Diagnosis.



Results

Preemptive Depression Identification Experiment

The results of the preemptive depression identification experiment are presented in [Tables 5-8](#). Each table shows a variation in the number of matched control users. Positive predictive value, sensitivity, and F_1 -score were used to measure the performance in each variation. The positive predictive value denotes the number of users classified as depressed who were indeed depressed. Sensitivity denotes how many of the depressed users were correctly classified as depressed. The F_1 -score, which is the harmonic mean of the positive predictive value and sensitivity, is suitable for use with data sets such as ours, where the class distribution (of depressed and controls) is uneven [69]. In contrast, accuracy is not suitable for such data sets [70]. Therefore, we used F_1 -score as the primary performance metric.

Using F_1 -score as a primary performance indicator, MentalBERT performs best across the variations.

A detailed breakdown of the results of the preemptive depression identification experiment can be found in [Multimedia Appendix 1](#).

Word embeddings (vector representations) result in strong sensitivity (recall), whereas TF-IDF features cause deficient performance. The positive predictive value (precision) was best observed when using the specialist LM, MentalBERT. The best F_1 -score was also achieved by MentalBERT and exceeded the naive baseline.

We now consider the selected users from RSDD-Matched and the performance of the classifiers against them. We will examine one misclassified user per variation in the experiment (in terms

of depressed users and the number of matched controls). For each variation, we will examine the strongest performing classifier and the user that it misclassified with the highest probability.

To identify the potential reasons for the misclassifications, we examined the lexical properties of user posts using 3 approaches. The first approach involves ascertaining the chief topic conveyed by the posts, a topic represented by 5 words. Topic modeling via latent Dirichlet allocation was used to accomplish this [71,72]. The second approach examines the chief TF-IDF features of the user posts. The third approach is to count the frequencies of depressed and control vocabularies ([Table 2](#)) that appear across the posts.

We present the misclassified depressed users with respect to each variation in the experiment ([Table 9](#)). We also present the misclassified control users with respect to each variation ([Table 10](#)).

One depressed user is often misclassified. User d13 was deemed a control user using 3 different classifiers across 3 different variations. Although depressed vocabulary counts slightly outweigh their control counterparts, the totals for both vocabularies were nominal. The topic of the user's posts is probably more indicative of the reasons for the misclassification. Certainly, a theme concerning death or dying appears to be present, but this is diluted by optimistic sounding references of temporal and geographic nature. Further diluting references are revealed among the TF-IDF features, where strong terms such as "love" are present. It seems that the classifiers construe such references as those belonging to a control user.

User d38 may have been misclassified for similar reasons. Counts for both depressed and control vocabularies were small.

Positive terms, such as “welcome” and “invite” might be deemed to belong to a control user.

An inferior performance was observed across the classifiers in the most imbalanced environment. We examine depressed user d57, which has been misclassified with a probability close to certainty. The depressed vocabulary count dwarfs the control vocabulary count. However, when making its decision, the classifier seems to harness the overarching nature of the user’s posts, as indicated by the topic model and TF-IDF features. The prevalence of “good” natured posts will inevitably see the user deemed similar to a control user when represented in a vector space.

We now consider misclassified control users with respect to each variation in the experiment (Table 10).

Certain users appear to be confounding across several different classifiers and variations. User c13 was strongly misclassified as a depressed user by both MentalBERT and MentalRoBERTa in the relatively noisy environments of 3 and 5 matched control users, respectively (Table 10). The depressed vocabulary counts far outweigh the control vocabulary counts for this user. In addition, the theological topic and TF-IDF features of the user’s

posts are deemed likely to be those of a depressed user, according to the classifier.

MentalBERT demonstrated adeptness in the most balanced variation in the experiment. We sought possible explanations for the misclassification of user c521. The control vocabulary count slightly outweighed that of depressed vocabulary. Moreover, the topic model and TF-IDF features are composed of terms that complement the control vocabulary. Intuitive reasons for misclassification as depressed are difficult to cite. Therefore, it is possible that, in a balanced environment, the classifier simply has too few control users to compare with depressed users.

In the noisiest environment, the simpler word-based model (SVM using word embeddings) demonstrated the strongest performance. Transformer-based language modeling cannot be performed. The vocabulary of the most strongly misclassified user in this case (c535) only offers a tenuous explanation. The count of depressed vocabulary was small, although it outweighed that of the control vocabulary. However, the topic and TF-IDF terms appeared to complement the depressed vocabulary, which may have been the cause of the misclassification.

Table 5. Binary classification scores using all posts of 70 depressed users and 1 of their matched control users^a.

	Positive predictive value, mean (SD)	Sensitivity, mean (SD)	F_1 -score, mean (SD)
SVM ^b using TF-IDF ^c	0.637 (N/A ^d)	0.557 (N/A)	0.590 (N/A)
SVM using word embeddings	0.558 (N/A)	0.543 (N/A)	0.548 (N/A)
SVM using TF-IDF and word embeddings	0.673 (N/A)	0.557 (N/A)	0.596 (N/A)
BERT ^e LM ^f	0.638 (0.021)	0.805 (0.022)	0.709 (0.012)
ALBERT ^g LM	0.606 (0.008)	0.786 (0.015)	0.683 (0.010)
BioBERT LM	0.601 (0.005)	0.862 (0.022)	0.707 (0.005)
Longformer LM	0.633 (0.009)	0.838 (0.036)	0.719 (0.018)
MentalBERT LM	0.660 (0.019)	0.848 (0.008)	0.738 (0.013)
MentalRoBERTa LM	0.629 (0.002)	0.819 (0.022)	0.709 (0.006)
Naive baseline—all depression	0.500 (N/A)	1.000 (N/A)	0.667 (N/A)

^aLanguage model experiments were run 3 times each, therefore both mean and SD scores are provided.

^bSVM: support vector machine.

^cTF-IDF: term frequency–inverse document frequency.

^dN/A: not applicable.

^eBERT: Bidirectional Encoder Representations from Transformers.

^fLM: language model.

^gALBERT: A Lite Bidirectional Encoder Representations from Transformers.

Table 6. Binary classification scores using all posts of 70 depressed users and 3 of their matched control users^a.

	Positive predictive value, mean (SD)	Sensitivity, mean (SD)	F_1 -score, mean (SD)
SVM ^b using TF-IDF ^c	0.800 (N/A ^d)	0.086 (N/A)	0.153 (N/A)
SVM using word embeddings	0.411 (N/A)	0.529 (N/A)	0.459 (N/A)
SVM using TF-IDF and word embeddings	0.800 (N/A)	0.057 (N/A)	0.107 (N/A)
BERT ^e LM ^f	0.653 (0.033)	0.481 (0.022)	0.546 (0.025)
ALBERT ^g LM	0.652 (0.034)	0.476 (0.009)	0.547 (0.018)
BioBERT LM	0.654 (0.028)	0.410 (0.030)	0.496 (0.020)
Longformer LM	0.653 (0.036)	0.476 (0.036)	0.534 (0.031)
MentalBERT LM	0.657 (0.034)	0.509 (0.008)	0.562 (0.016)
MentalRoBERTa LM	0.614 (0.023)	0.471 (0.015)	0.522 (0.002)
Naive baseline—all depression	0.250 (N/A)	1.000 (N/A)	0.167 (N/A)

^aLanguage model experiments were run 3 times each, therefore both mean and SD scores are provided.

^bSVM: support vector machine.

^cTF-IDF: term frequency–inverse document frequency.

^dN/A: not applicable.

^eBERT: Bidirectional Encoder Representations from Transformers.

^fLM: language model.

^gALBERT: A Lite Bidirectional Encoder Representations from Transformers.

Table 7. Binary classification scores using all posts of 70 depressed users and 5 of their matched control users^a.

	Positive predictive value, mean (SD)	Sensitivity, mean (SD)	F_1 -score, mean (SD)
SVM ^b using TF-IDF ^c	0.400 (N/A ^d)	0.029 (N/A)	0.053 (N/A)
SVM using word embeddings	0.309 (N/A)	0.471 (N/A)	0.372 (N/A)
SVM using TF-IDF and word embeddings	0.200 (N/A)	0.014 (N/A)	0.027 (N/A)
BERT ^e LM ^f	0.615 (0.028)	0.290 (0.022)	0.379 (0.017)
ALBERT ^g LM	0.555 (0.030)	0.281 (0.009)	0.354 (0.006)
BioBERT LM	0.627 (0.034)	0.252 (0.021)	0.331 (0.027)
Longformer LM	0.624 (0.108)	0.286 (0.038)	0.363 (0.059)
MentalBERT LM	0.572 (0.002)	0.329 (0.043)	0.400 (0.040)
MentalRoBERTa LM	0.562 (0.027)	0.343 (0.000)	0.419 (0.010)
Naive baseline—all depression	0.167 (N/A)	1.000 (N/A)	0.286 (N/A)

^aLanguage model experiments were run 3 times each, therefore both mean and SD scores are provided.

^bSVM: support vector machine.

^cTF-IDF: term frequency–inverse document frequency.

^dN/A: not applicable.

^eBERT: Bidirectional Encoder Representations from Transformers.

^fLM: language model.

^gALBERT: A Lite Bidirectional Encoder Representations from Transformers.

Table 8. Binary classification scores using all posts of 70 depressed users and 10 of their matched control users^a.

	Positive predictive value, mean (SD)	Sensitivity, mean (SD)	F_1 -score, mean (SD)
SVM ^b using TF-IDF ^c	0.000 (N/A) ^d	0.000 (N/A)	0.000 (N/A)
SVM using word embeddings	0.212 (N/A)	0.371 (N/A)	0.268 (N/A)
SVM using TF-IDF and word embeddings	0.000 (N/A)	0.000 (N/A)	0.000 (N/A)
BERT ^e LM ^f	0.100 (0.000)	0.014 (0.000)	0.025 (0.00)
ALBERT ^g LM	0.089 (0.019)	0.014 (0.000)	0.025 (0.001)
BioBERT LM	0.067 (0.115)	0.005 (0.008)	0.009 (0.016)
Longformer LM	0.024 (0.019)	0.019 (0.033)	0.021 (0.037)
MentalBERT LM	0.167 (0.058)	0.014 (0.000)	0.026 (0.001)
MentalRoBERTa LM	0.272 (0.185)	0.034 (0.008)	0.057 (0.018)
Naive baseline—all depression	0.091 (N/A)	1.000 (N/A)	0.167 (N/A)

^aLanguage model experiments were run 3 times each, therefore both mean and SD scores are provided.

^bSVM: support vector machine.

^cTF-IDF: term frequency–inverse document frequency.

^dN/A: not applicable.

^eBERT: Bidirectional Encoder Representations from Transformers.

^fLM: language model.

^gALBERT: A Lite Bidirectional Encoder Representations from Transformers.

Table 9. Depressed users most strongly misclassified in each variation of the preemptive depression identification experiment^a.

	One depression user per control user (1:1)	One depression user per 3 control users (1:3)	One depression user per 5 control users (1:5)	One depression user per 10 control users (1:10)
Classifier	MentalBERT LM ^b	MentalBERT LM	MentalRoBERTa LM	SVM ^c using word embeddings
User	d13	d38	d13	d57
Control probability	0.93	0.94	0.99	0.98
Sum of post lengths in words	1696	1888	1696	55,897
Topic	<ul style="list-style-type: none"> • news • hawaii • time • dead • blue 	<ul style="list-style-type: none"> • sir-geo • welcomed • invite • leave • warlock 	<ul style="list-style-type: none"> • news • hawaii • time • dead • blue 	<ul style="list-style-type: none"> • good • time • people • years • problem
Chief TF-IDF ^d features	<ul style="list-style-type: none"> • love • minnesota • diablo • time • man • bud • zoidberg • like • month • hawaii 	<ul style="list-style-type: none"> • sir • geo • welcome • invite • warlock • leave • titan • psn • run • need 	<ul style="list-style-type: none"> • love • minnesota • diablo • time • man • bud • zoidberg • like • month • hawaii 	<ul style="list-style-type: none"> • good • know • use • make • time • thank • link • want • try • like

Depressed vocabulary counts

people	1	1	1	64
know	6	0	6	93
thing	3	0	3	35
feel	2	2	2	10
time	5	8	5	99
woman	1	0	1	7
go	3	0	3	54
want	3	1	3	71
life	2	0	2	28
relationship	0	0	0	2

Control vocabulary counts

game	0	1	0	9
trade	0	0	0	2
key	0	0	0	4
team	2	3	2	4
play	0	1	0	35
player	0	0	0	8
shiny	0	0	0	0
hatch	0	0	0	0
thank	1	1	0	15
add	0	2	0	14

^aLexical properties of those users' posts are provided.^bLM: language model.^cSVM: support vector machine.^dTF-IDF: term frequency–inverse document frequency.

Table 10. Control users most strongly misclassified in each variation of the preemptive depression identification experiment^a.

	One depression user per control user (1:1)	One depression user per 3 control users (1:3)	One depression user per 5 control users (1:5)	One depression user per 10 control users (1:10)
Classifier	MentalBERT LM ^b	MentalBERT LM	MentalRoBERTa LM	SVM ^c using Word embeddings
User	c521	c13	c13	c535
Depressed probability	0.99	0.95	0.91	0.91
Sum of post lengths in words	1513	8489	8489	1595
Topic	<ul style="list-style-type: none"> • elo • play • team • bronze • games 	<ul style="list-style-type: none"> • god • jesus • people • good • life 	<ul style="list-style-type: none"> • god • jesus • people • good • life 	<ul style="list-style-type: none"> • people • shit • reddit • guy • man
Chief TF-IDF ^d features	<ul style="list-style-type: none"> • team • just • suck • elo • play • game • like • good • sydtko • win 	<ul style="list-style-type: none"> • god • think • way • thing • try • know • jesus • people • say • like 	<ul style="list-style-type: none"> • god • think • way • thing • try • know • jesus • people • say • like 	<ul style="list-style-type: none"> • say • thank • guy • people • reddit • man • make • tell • watch • let
Depressed vocabulary counts				
people	4	48	48	6
know	2	36	36	3
thing	3	28	28	1
feel	1	6	6	1
time	2	6	6	4
woman	0	4	4	0
go	0	4	4	5
want	3	16	16	1
life	0	46	46	1
relationship	0	8	8	0
Control vocabulary counts				
game	7	0	0	0
trade	0	0	0	0
key	0	0	0	0
team	9	0	0	0
play	9	6	6	0
player	2	0	0	0
shiny	0	0	0	0
hatch	0	0	0	0
thank	1	4	4	1
add	1	0	0	0

^aLexical properties of those users' posts are provided.

^bLM: language model.

^cSVM: Support Vector Machine.

^dTF-IDF: Term Frequency—Inverse Document Frequency.

Temporal Experiment

We then performed a temporal experiment. Because BERT achieved the highest F_1 -score across all preemptive depression identification experiment variations, it was selected as the exclusive general-purpose LM here. For the same reason, MentalBERT was selected as an exclusive specialist LM. The results are presented in [Tables 11](#) and [12](#). Each table shows a variation in the number of matched control users. The average performance of each LM across the 2 variations is shown in [Figure 2](#).

For BERT, the strongest sensitivity and F_1 -scores were observed when only 12 weeks (approximately 3 months) of posts before the estimated diagnosis dates were considered. Subsets larger or smaller than 12 weeks caused degradation in the classifier performance. For MentalBERT, the strongest sensitivity and F_1 -scores were obtained when either 16 or 24 weeks of posts were considered. With BERT scoring a higher F_1 -score at 12 weeks than MentalBERT, this suggests that the final 12 weeks of posts before a depressed user's estimated diagnosis date may be the most indicative of their illness.

An explanation for the slightly inferior performance of MentalBERT may be found in its construction: it is pretrained on text from mental health subreddits such as “r/depression” and “r/mental health” [48]. However, RSDD (from which we derived RSDD-Matched) does not contain posts from mental health subreddits. Therefore, when RSDD-Matched data are limited, as in our temporal experiment, more general-purpose models, such as BERT, may be able to achieve stronger performance. BERT is pretrained on more general corpora, such as Wikipedia [44].

A detailed breakdown of the results of the temporal experiment can be found in [Multimedia Appendix 1](#).

We once again consider selected users from RSDD-Matched and the performance of the classifiers against them. We again examined one misclassified user per variation in the experiment (in terms of depressed users and number of matched controls). For each variation, we will examine the strongest performing

time span, and the user that is misclassified with the highest probability. To identify the reasons for the misclassifications, we again examined the lexical properties of the user posts using topic models, TF-IDF features, and vocabulary ([Table 2](#)) frequency counts.

Misclassified depressed users with respect to the 2 variations in the experiment are listed in [Table 13](#).

User d52 is a depressed user misclassified in both balanced and imbalanced environments, where only the final 12 weeks of their posts are considered. The vocabulary of these posts intersected with very little of the chief depressed vocabulary. It intersects with slightly more of the chief control vocabulary. The topic and TF-IDF features, intuitively speaking, appear to belong to that of a control rather than a depressed user. Perhaps, a balanced environment with temporally limited post histories provides little training data from which the classifier can learn to differentiate between controls and depressed users. Although rare, these cases may occur in practice and highlight the importance of being careful in overrelying on automatic models for individual assessments without human expert intervention.

We now consider the misclassified control users with respect to the 2 variations in the experiment ([Table 14](#)).

First, we consider user c481. Both its depressed and control vocabulary counts were zero, which offers some insight into misclassification. The topic and TF-IDF features of the posts appear to align with those of the control user. However, it is likely that the prevalence of “pain” is a confounding factor. This term may be intuitively linked to depressed users, which may mislead the classifier. Again, the limited temporal range of posts in this setting provided little data from which the classifier could learn.

User c13 is a confounder in the preemptive depression identification experiment and has been proven to be so in the temporal experiment. Even when considering only the last 12 weeks of the user's posts in an imbalanced environment, theologically themed vocabulary is not diluted. It intersects strongly with the vocabulary of depressed users and explains this misclassification.

Table 11. Binary classification scores using 56 depressed users and 1 of their matched control users and 6 temporal post subsets^a.

	Positive predictive value, mean (SD)	Sensitivity, mean (SD)	F_1 -score, mean (SD)
Last 4 weeks			
BERT ^b LM ^c	0.575 (0.027)	0.830 (0.039)	0.675 (0.023)
MentalBERT LM	0.612 (0.026)	0.835 (0.026)	0.698 (0.017)
Last 8 weeks			
BERT LM	0.598 (0.026)	0.854 (0.071)	0.700 (0.037)
MentalBERT LM	0.603 (0.020)	0.842 (0.047)	0.699 (0.022)
Last 12 weeks			
BERT LM	0.605 (0.014)	0.912 (0.018)	0.726 (0.015)
MentalBERT LM	0.600 (0.013)	0.888 (0.010)	0.715 (0.008)
Last 16 weeks			
BERT LM	0.570 (0.009)	0.863 (0.026)	0.684 (0.007)
MentalBERT LM	0.575 (0.009)	0.907 (0.028)	0.703 (0.016)
Last 20 weeks			
BERT LM	0.569 (0.023)	0.893 (0.036)	0.694 (0.025)
MentalBERT LM	0.578 (0.018)	0.882 (0.027)	0.696 (0.014)
Last 24 weeks			
BERT LM	0.565 (0.021)	0.871 (0.027)	0.683 (0.010)
MentalBERT LM	0.591 (0.014)	0.890 (0.010)	0.707 (0.011)
All posts			
BERT LM	0.627 (0.018)	0.824 (0.032)	0.710 (0.019)
MentalBERT LM	0.638 (0.009)	0.861 (0.000)	0.732 (0.006)
Naive baseline	0.500 (N/A ^d)	1.000 (N/A)	0.667 (N/A)

^aThe classifiers used are BERT LM and MentalBERT LM, both of whose experiments were run 3 times each, therefore both mean and SD scores are provided.

^bBERT: Bidirectional Encoder Representations From Transformers.

^cLM: language model.

^dN/A: not applicable.

Table 12. Binary classification scores using 56 depressed users and 3 of their matched control users and 6 temporal post subsets^a.

	Positive predictive value, mean (SD)	Sensitivity, mean (SD)	F_1 -score, mean (SD)
Last 4 weeks			
BERT ^b LM ^c	0.480 (0.027)	0.538 (0.019)	0.489 (0.010)
MentalBERT LM	0.494 (0.019)	0.577 (0.009)	0.525 (0.007)
Last 8 weeks			
BERT LM	0.446 (0.032)	0.538 (0.036)	0.472 (0.035)
MentalBERT LM	0.427 (0.027)	0.524 (0.029)	0.461 (0.023)
Last 12 weeks			
BERT LM	0.498 (0.031)	0.619 (0.037)	0.543 (0.035)
MentalBERT LM	0.448 (0.007)	0.569 (0.017)	0.494 (0.009)
Last 16 weeks			
BERT LM	0.471 (0.010)	0.565 (0.021)	0.504 (0.011)
MentalBERT LM	0.481 (0.023)	0.643 (0.037)	0.541 (0.028)
Last 20 weeks			
BERT LM	0.475 (0.039)	0.577 (0.037)	0.510 (0.034)
MentalBERT LM	0.487 (0.018)	0.595 (0.011)	0.524 (0.009)
Last 24 weeks			
BERT LM	0.470 (0.033)	0.591 (0.036)	0.518 (0.033)
MentalBERT LM	0.501 (0.022)	0.591 (0.018)	0.536 (0.022)
All posts			
BERT LM	0.625 (0.021)	0.519 (0.032)	0.562 (0.015)
MentalBERT LM	0.588 (0.005)	0.508 (0.010)	0.540 (0.003)
Naive baseline	0.250 (N/A ^d)	1.000 (N/A)	0.400 (N/A)

^aThe classifiers used are BERT LM and MentalBERT LM, both of whose experiments were run 3 times each, therefore both mean and SD scores are provided..

^bBERT: Bidirectional Encoder Representations From Transformer.

^cLM: language model.

^dN/A: not applicable.

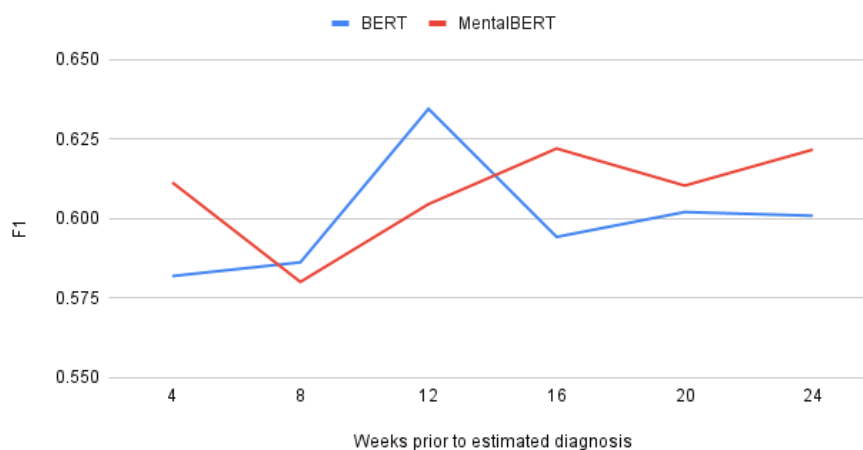
Figure 2. Average performances of Bidirectional Encoder Representations from Transformers (BERT) and MentalBERT between 4 and 24 weeks before the estimated diagnosis date.

Table 13. Depressed users most strongly misclassified in each variation of the temporal experiment. Lexical properties of those users' posts are provided.

	One depression user per control user (1:1)	One depression user per 3 control users (1:3)
Time span	Last 12 weeks	Last 12 weeks
Classifier	BERT ^a LM ^b	BERT LM
User	d52	d52
Control probability	0.869	0.935
Sum of post lengths in words	1225	1225
Topic	england belgium hamster time team	england belgium hamster time team
Chief TF-IDF ^c features	<ul style="list-style-type: none"> • thank • team • player • help • time • goal • cage • post • second • start 	<ul style="list-style-type: none"> • thank • team • player • help • time • goal • cage • post • second • start
Depressed vocabulary counts		
people	0	0
know	1	1
thing	1	1
feel	0	0
time	4	4
woman	0	0
go	0	0
want	2	2
life	0	0
relationship	0	0
Control vocabulary counts		
game	2	2
trade	0	0
key	0	0
team	4	4
play	0	0
player	1	1
shiny	0	0
hatch	0	0
thank	2	2
add	1	1

^aBERT: Bidirectional Encoder Representations From Transformers.

^bLM: language model.

^cTF-IDF: term frequency–inverse document frequency.

Table 14. Control users most strongly misclassified in each variation of the temporal experiment. Lexical properties of those users' posts are provided.

	One depression user per control user (1:1)	One depression user per 3 control users (1:3)
Time span	Last 12 weeks	Last 12 weeks
Classifier	BERT ^a LM ^b	BERT LM
User	c481	c13
Depressed probability	0.963	0.917
Total length of posts in words	258	8489
Topic	<ul style="list-style-type: none"> • food • clove • tomorrow • pain • suspect 	<ul style="list-style-type: none"> • god • jesus • people • good • life
Chief TF-IDF ^c features	<ul style="list-style-type: none"> • reply • eat • food • cat • clove • pain • suspect • tooth • vet • water 	<ul style="list-style-type: none"> • god • think • way • thing • try • know • jesus • people • say • like
Depressed vocabulary counts		
people	0	24
know	0	18
thing	0	14
feel	0	3
time	0	3
woman	0	2
go	0	2
want	0	8
life	0	23
relationship	0	4
Control vocabulary counts		
game	0	0
trade	0	0
key	0	0
team	0	0
play	0	3
player	0	0
shiny	0	0
hatch	0	0
thank	0	2
add	0	0

^aBERT: Bidirectional Encoder Representations From Transformers.

^bLM: language model.

^cTF-IDF: term frequency–inverse document frequency.

Sentiment Analysis

A sentiment analysis was then performed to complement the temporal experiment. We present the band-wise changes in sentiment for each class (Figures 3 and 4). It is observed that negatively charged posts for depressed users are less frequent as we approach the (estimated) diagnosis date, which may be deemed counterintuitive (Figure 3). However, it is also notable that depressed users' posts were, on average, more negative than those of control users throughout the 24-week period (Figure 4). This aligns with previous studies that found a positive correlation between mental illness and negative sentiments [73].

We then sought to establish whether the diagnosis was associated with the sentiment of the post. The results of the logistic regression model (Table 15) indicate that there is a clear significant association between the diagnosis and the

“sentimentality” of the post ($P < .05$), despite no apparent effect of temporality. Interestingly, the word count of a post appeared as a significant covariate of this model ($P = .001$), indicating that longer posts are slightly more likely to be classified as “sentimental,” irrespective of the depression status of the user.

Table 16 presents the results of the Multinomial Regression Model. Again, all effect size estimates were compatible with our inferences on the basis of a simpler logistic model. However, the multinomial analysis gives us an additional perspective: the effects of depression diagnosis are similar between positive and negative sentiments, with overlapping CIs statistically indistinguishable. This is the case despite the varying effects of other covariates, such as word count, which displays regression β coefficients of opposite signs in both sentiments (more words associate with negative posts, whereas fewer words associate with positive posts).

Figure 3. Change in the average percentage of positive and negative posts across 6 temporal bands: 0 to 4, 4 to 8, 8 to 12, 12 to 16, 16 to 20, and 20 to 24 weeks before the estimated diagnosis date (for a control user, this is the estimated diagnosis of its matched depressed user).

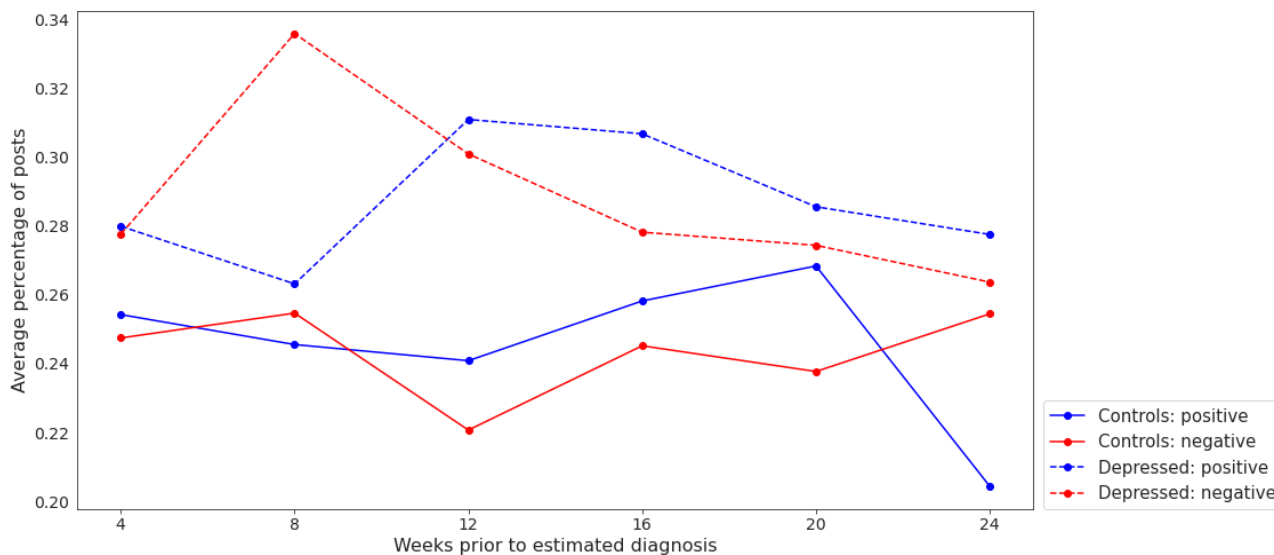


Figure 4. Average percentage of positive and negative posts per temporal band. Temporal bands include 0 to 4, 4 to 8, 8 to 12, 12 to 16, 16 to 20, and 20 to 24 weeks before the estimated diagnosis date (for a control user, this is the estimated diagnosis of its matched depressed user).

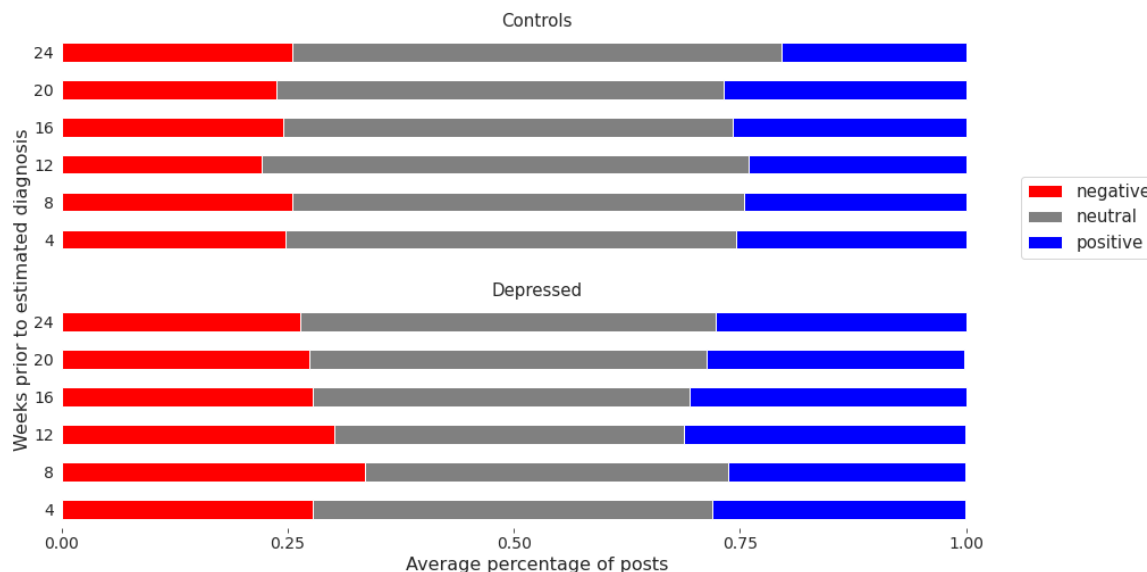


Table 15. Logistic regression results for predicting whether a post is neutral or not neutral.

Variable	β	Odds ratio	SE	<i>P</i> value
Depression diagnosis	0.163	1.177	0.035	<.001
Time to diagnosis	-0.004	0.996	0.013	.75
Post word count	0.040	1.041	0.012	.001
Interaction (diagnosis \times time)	0.011	1.011	0.013	.41

Table 16. Multinomial regression results for predicting whether a post is positive or negative.

Sentiment and variable	β	Odds ratio	SE	<i>P</i> value
Positive				
Depression diagnosis	0.190	1.209	0.047	<.001
Time to diagnosis	0.015	1.015	0.016	.37
Post word count	-0.070	0.932	0.019	<.001
Interaction (diagnosis \times time)	0.045	1.046	0.016	.006
Negative				
Depression diagnosis	0.151	1.163	0.041	<.001
Time to diagnosis	-0.019	0.981	0.016	.24
Post word count	0.103	1.108	0.014	<.001
Interaction (diagnosis \times time)	-0.021	0.979	0.016	.18

Discussion

Principal Findings

We obtained evidence that LMs (particularly BERT-like models) can be used in preemptive mental health detection and analysis in longhand forums, even if they have room for improvement.

In our preemptive depression detection experiment, depressed and control subjects were placed in ratios of 1:1, 1:3, 1:5, and 1:10. The purpose was to simulate increasingly realistic settings in which most users were controls. In the balanced arrangement of 1:1, we obtained an F_1 -score of 0.738 using the MentalBERT LM. This is comparable with the works of Eichstaedt et al [14], de Choudhury et al [74], and Reece et al [19], who obtained F_1 -scores of 0.660, 0.680, and 0.650, respectively. This study provides evidence that LMs are more effective than existing methods for predicting depression in social media data before diagnosis.

Our temporal analysis suggested that the final 12 weeks (approximately 3 months) of posts before a depressed user's estimated diagnosis date are likely to be the most indicative of their condition. Another broader interpretation is that LMs do not appear to improve with the addition of more data before 12-16 weeks. The BERT and MentalBERT obtained F_1 -scores of 0.726 and 0.715, respectively.

This is in contrast to a certain extent with the results of Eichstaedt et al [14], albeit using area under curve scores rather than F_1 -scores. Six months before the diagnosis date, 0.72 was obtained, and 3 months prior, 0.62 was obtained. From these results, it is difficult to draw clear conclusions because the

results may be affected by the nature of the data and models used.

We also observed that posts made during the 4- to 8-week period before the user's estimated diagnosis date are also pertinent. They exhibited more negative sentiment than posts made during any other 4-week period (up to 24 weeks before their estimated diagnosis date). This finding may be supportive of prior work that distinct changes in mood may be predictive of the onset of depression [75].

We were able to corroborate the importance of sentiment in the discourse of depressed users. We found that depressed users are approximately 1.18 times more likely to make a sentimental post than nondepressed users.

Limitations

Constraints on our investigation primarily concern RSDD-Matched, where 70 depressed users make up a small sample. However, use 5-fold cross-validation to mitigate this and performed different experiments with various numbers of control users.

RSDD-Matched is derived from RSDD and RSDD-Time. As a result, the diagnosis dates of the users in RSDD-Matched are estimates only. Furthermore, posts made in mental health subreddits were deliberately elided from the RSDD and were not available for consideration by our machine classifiers.

Conclusions

Using state-of-the-art LMs, this study posits how far the diagnosis of depression in a person with depressive traits can be determined in advance. With this knowledge, it may be possible to direct people with depression to physicians much

sooner than they would otherwise. Moreover, perhaps more importantly, we have shown how these automatic NLP tools can serve to analyze the main traits arising from web-based posts.

We have also observed that the sentiment exhibited in web-based forum postings demonstrates good sensitivity in detecting depressive traits.

Further work may include a multimodal approach to the detection of people with depression in web-based forums such as Reddit. For example, along with the text of Reddit users' posts, we might also consider the subreddits where they have

upvoted and downvoted posts. The awards received or given may also indicate a user's mental health. Such a study would, of course, be contingent on the ability to synthesize a suitable data set or source an existing one. Moreover, the use of temporal information such as temporal word embeddings [76] may enhance any multimodal approach.

Methods for gauging the severity of depression in web-based forum users should also be investigated. This might involve mining language features from user posts and observing how they correlate with ground-truth severity. Features of interest may include terms used in Linguistic Inquiry and Word Count dictionaries, sentiment, and emotion [77].

Acknowledgments

AFP was supported by the Academy of Medical Sciences "Springboard" award (SBF005 \ 1083). JCC is supported by a UK Research and Innovation (UKRI) Future Leaders Fellowship. The authors thank Professor Nazli Goharian of Georgetown University and Dr Andrew Yates of University of Amsterdam for their assistance in supplying Reddit Self-reported Depression Diagnosis (RSDD) and RSDD-Time.

Data Availability

Information on the RSDD and RSDD-Time data sets used in this study, including their data access procedure, can be found on the web [78].

Conflicts of Interest

None declared.

Multimedia Appendix 1

Reddit Self-reported Depression Diagnosis—Matched metadata and verbose results of the preemptive and temporal experiments. [[XLSX File \(Microsoft Excel File\), 956 KB](#) - [ai_v21e41205_app1.xlsx](#)]

References

1. Global Health Data Exchange (GHDx). Institute of Health Metrics and Evaluation. URL: <http://ghdx.healthdata.org/gbd-results-tool?params=gbd-api-2019-permalink/d780dffbe8a381b25e1416884959e88b> [accessed 2021-05-01]
2. Kessler RC, Petukhova M, Sampson NA, Zaslavsky AM, Wittchen H. Twelve-month and lifetime prevalence and lifetime morbid risk of anxiety and mood disorders in the United States. *Int J Methods Psychiatr Res* 2012 Sep;21(3):169-184 [[FREE Full text](#)] [doi: [10.1002/mpr.1359](https://doi.org/10.1002/mpr.1359)] [Medline: [22865617](https://pubmed.ncbi.nlm.nih.gov/22865617/)]
3. Regier DA, Kuhl EA, Kupfer DJ. The DSM-5: classification and criteria changes. *World Psychiatry* 2013 Jun 04;12(2):92-98 [[FREE Full text](#)] [doi: [10.1002/wps.20050](https://doi.org/10.1002/wps.20050)] [Medline: [23737408](https://pubmed.ncbi.nlm.nih.gov/23737408/)]
4. Picardi A, Lega I, Tarsitani L, Caredda M, Matteucci G, Zerella M, SET-DEP Group. A randomised controlled trial of the effectiveness of a program for early detection and treatment of depression in primary care. *J Affect Disord* 2016 Jul 01;198:96-101. [doi: [10.1016/j.jad.2016.03.025](https://doi.org/10.1016/j.jad.2016.03.025)] [Medline: [27015158](https://pubmed.ncbi.nlm.nih.gov/27015158/)]
5. Edwards S, Tinning L, Brown JS, Boardman J, Weinman J. Reluctance to seek help and the perception of anxiety and depression in the United Kingdom: a pilot vignette study. *J Nerv Ment Dis* 2007 Mar;195(3):258-261. [doi: [10.1097/01.nmd.0000253781.49079.53](https://doi.org/10.1097/01.nmd.0000253781.49079.53)] [Medline: [17468687](https://pubmed.ncbi.nlm.nih.gov/17468687/)]
6. Wasserman C, Hoven CW, Wasserman D, Carli V, Sarchiapone M, Al-Halabi S, et al. Suicide prevention for youth--a mental health awareness program: lessons learned from the Saving and Empowering Young Lives in Europe (SEYLE) intervention study. *BMC Public Health* 2012 Sep 12;12:776 [[FREE Full text](#)] [doi: [10.1186/1471-2458-12-776](https://doi.org/10.1186/1471-2458-12-776)] [Medline: [22971152](https://pubmed.ncbi.nlm.nih.gov/22971152/)]
7. De Choudhury M, Counts S, Horvitz E. Social media as a measurement tool of depression in populations. In: Proceedings of the 5th Annual ACM Web Science Conference. 2013 Presented at: WebSci '13: Web Science 2013; May 2 - 4, 2013; Paris France. [doi: [10.1145/2464464.2464480](https://doi.org/10.1145/2464464.2464480)]
8. Tadesse MM, Lin H, Xu B, Yang L. Detection of depression-related posts in Reddit social media forum. *IEEE Access* 2019;7:44883-44893. [doi: [10.1109/access.2019.2909180](https://doi.org/10.1109/access.2019.2909180)]
9. Malhotra A, Jindal R. Deep learning techniques for suicide and depression detection from online social media: a scoping review. *Applied Soft Computing* 2022 Nov;130:109713. [doi: [10.1016/j.asoc.2022.109713](https://doi.org/10.1016/j.asoc.2022.109713)]

10. Zhang T, Schoene AM, Ji S, Ananiadou S. Natural language processing applied to mental illness detection: a narrative review. *NPJ Digit Med* 2022 Apr 08;5(1):46 [FREE Full text] [doi: [10.1038/s41746-022-00589-7](https://doi.org/10.1038/s41746-022-00589-7)] [Medline: [35396451](https://pubmed.ncbi.nlm.nih.gov/35396451/)]
11. Yates A, Cohan A, Goharian N. Depression and self-harm risk assessment in online forums. In: *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*. 2017 Presented at: 2017 Conference on Empirical Methods in Natural Language Processing; Sep 7–11, 2017; Copenhagen, Denmark. [doi: [10.18653/v1/d17-1322](https://doi.org/10.18653/v1/d17-1322)]
12. MacAvaney S, Desmet B, Cohan A, Soldaini L, Yates A, Zirikly A, et al. RSDD-time: temporal annotation of self-reported mental health diagnoses. In: *Proceedings of the Fifth Workshop on Computational Linguistics and Clinical Psychology: From Keyboard to Clinic*. 2018 Presented at: Fifth Workshop on Computational Linguistics and Clinical Psychology: From Keyboard to Clinic; Jun 5, 2018; New Orleans, LA. [doi: [10.18653/v1/w18-0618](https://doi.org/10.18653/v1/w18-0618)]
13. Owen D, Camacho-Collados J, Anke L. Towards preemptive detection of depression and anxiety in Twitter. *arXiv* 2020 Nov [FREE Full text]
14. Abed-Esfahani P, Howard D, Maslej M, Patel S, Mann V, Goegan S, et al. Transfer learning for depression: early detection and severity prediction from social media postings. *CAMH*. 2019. URL: https://ceur-ws.org/Vol-2380/paper_102.pdf [accessed 2022-03-05]
15. Shah F, Ahmed F, Joy S, Ahmed S, Sadek S, Shil R, et al. Early depression detection from social network using deep learning techniques. In: *Proceedings of the IEEE Region 10 Symposium (TENSYP)*. 2020 Presented at: IEEE Region 10 Symposium (TENSYP); Jun 05-07, 2020; Dhaka, Bangladesh. [doi: [10.1109/tensymp50017.2020.9231008](https://doi.org/10.1109/tensymp50017.2020.9231008)]
16. Eichstaedt JC, Smith RJ, Merchant RM, Ungar LH, Crutchley P, Preoțiu-Pietro D, et al. Facebook language predicts depression in medical records. *Proc Natl Acad Sci U S A* 2018 Oct 30;115(44):11203-11208 [FREE Full text] [doi: [10.1073/pnas.1802331115](https://doi.org/10.1073/pnas.1802331115)] [Medline: [30322910](https://pubmed.ncbi.nlm.nih.gov/30322910/)]
17. Ren L, Lin H, Xu B, Zhang S, Yang L, Sun S. Depression detection on reddit with an emotion-based attention network: algorithm development and validation. *JMIR Med Inform* 2021 Jul 16;9(7):e28754 [FREE Full text] [doi: [10.2196/28754](https://doi.org/10.2196/28754)] [Medline: [34269683](https://pubmed.ncbi.nlm.nih.gov/34269683/)]
18. Winokur G. Duration of illness prior to hospitalization (onset) in the affective disorders. *Neuropsychobiology* 1976;2(2-3):87-93. [doi: [10.1159/000117535](https://doi.org/10.1159/000117535)] [Medline: [1012452](https://pubmed.ncbi.nlm.nih.gov/1012452/)]
19. Reece AG, Reagan AJ, Lix KL, Dodds PS, Danforth CM, Langer EJ. Forecasting the onset and course of mental illness with Twitter data. *Sci Rep* 2017 Oct 11;7(1):13006 [FREE Full text] [doi: [10.1038/s41598-017-12961-9](https://doi.org/10.1038/s41598-017-12961-9)] [Medline: [29021528](https://pubmed.ncbi.nlm.nih.gov/29021528/)]
20. van Beljouw IM, Verhaak PF, Cuijpers P, van Marwijk HW, Penninx BW. The course of untreated anxiety and depression, and determinants of poor one-year outcome: a one-year cohort study. *BMC Psychiatry* 2010 Oct 20;10:86 [FREE Full text] [doi: [10.1186/1471-244X-10-86](https://doi.org/10.1186/1471-244X-10-86)] [Medline: [20961414](https://pubmed.ncbi.nlm.nih.gov/20961414/)]
21. Ghio L, Gotelli S, Marcenaro M, Amore M, Natta W. Duration of untreated illness and outcomes in unipolar depression: a systematic review and meta-analysis. *J Affect Disord* 2014 Jan;152-154:45-51. [doi: [10.1016/j.jad.2013.10.002](https://doi.org/10.1016/j.jad.2013.10.002)] [Medline: [24183486](https://pubmed.ncbi.nlm.nih.gov/24183486/)]
22. Agorastos A, Marmar CR, Otte C. Immediate and early behavioral interventions for the prevention of acute and posttraumatic stress disorder. *Curr Opin Psychiatry* 2011 Nov;24(6):526-532. [doi: [10.1097/YCO.0b013e32834cdde2](https://doi.org/10.1097/YCO.0b013e32834cdde2)] [Medline: [21941180](https://pubmed.ncbi.nlm.nih.gov/21941180/)]
23. Guntuku SC, Yaden DB, Kern ML, Ungar LH, Eichstaedt JC. Detecting depression and mental illness on social media: an integrative review. *Curr Opinion Behavioral Sci* 2017 Dec;18:43-49. [doi: [10.1016/j.cobeha.2017.07.005](https://doi.org/10.1016/j.cobeha.2017.07.005)]
24. SMHD, RSDD, and RSDD-Time Datasets. Georgetown information retrieval lab. URL: https://docs.google.com/forms/d/e/1FAIpQLScC-O3MXDd2lZSGqERHsv1EMVR2xN5WC0cAodsHK3tBOz_FLw/viewform [accessed 2020-11-21]
25. Zhang Y, Jin R, Zhou Z. Understanding bag-of-words model: a statistical framework. *Int J Mach Learn Cyber* 2010 Aug 28;1(1-4):43-52. [doi: [10.1007/s13042-010-0001-0](https://doi.org/10.1007/s13042-010-0001-0)]
26. Read J. Recognising affect in text using pointwise-mutual information. University of Sussex. 2004 Sep. URL: <https://citeseerx.ist.psu.edu/document?repid=rep1&type=pdf&doi=11185d20f109d28295f4f4ec8a72f33023709137> [accessed 2022-03-31]
27. Salton G, Buckley C. Term-weighting approaches in automatic text retrieval. *Inform Process Manage* 1988 Jan;24(5):513-523. [doi: [10.1016/0306-4573\(88\)90021-0](https://doi.org/10.1016/0306-4573(88)90021-0)]
28. Lafon P. Sur la variabilité de la fréquence des formes dans un corpus. *Mots. Les langages du politique* 1980;1(1):127-165. [doi: [10.3406/mots.1980.1008](https://doi.org/10.3406/mots.1980.1008)]
29. Drouin P. Term extraction using non-technical corpora as a point of leverage. *Terminology* 2003 Sep 2;9(1):99-115. [doi: [10.1075/term.9.1.06dro](https://doi.org/10.1075/term.9.1.06dro)]
30. Camacho-Collados J, Pilehvar MT, Navigli R. Nasari: integrating explicit knowledge and corpus statistics for a multilingual representation of concepts and entities. *Artif Intell* 2016 Nov;240:36-64. [doi: [10.1016/j.artint.2016.07.005](https://doi.org/10.1016/j.artint.2016.07.005)]
31. Boser B, Guyon I, Vapnik V. A training algorithm for optimal margin classifiers. In: *Proceedings of the fifth annual workshop on Computational learning theory*. 1992 Presented at: COLT92: 5th Annual Workshop on Computational Learning Theory; Jul 27 - 29, 1992; Pittsburgh Pennsylvania USA URL: <https://dl.acm.org/doi/proceedings/10.1145/130385> [doi: [10.1145/130385.130401](https://doi.org/10.1145/130385.130401)]
32. Pirina I, Çöltekin C. Identifying depression on Reddit: the effect of training data. In: *Proceedings of the 2018 EMNLP Workshop SMM4H: The 3rd Social Media Mining for Health Applications Workshop & Shared Task*. 2018 Presented at:

- 2018 EMNLP Workshop SMM4H: The 3rd Social Media Mining for Health Applications Workshop & Shared Task; Oct, 2018; Brussels, Belgium. [doi: [10.18653/v1/w18-5903](https://doi.org/10.18653/v1/w18-5903)]
33. Salton G, Wong A, Yang CS. A vector space model for automatic indexing. *Commun ACM* 1975 Nov;18(11):613-620. [doi: [10.1145/361219.361220](https://doi.org/10.1145/361219.361220)]
34. Turian J, Ratnoff L, Bengio Y. Word representations: a simple and general method for semi-supervised learning. In: *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*. 2010 Presented at: 48th Annual Meeting of the Association for Computational Linguistics; Jul 11 - 16, 2010; Uppsala Sweden URL: <https://aclanthology.org/P10-1040.pdf>
35. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, et al. Scikit-learn: machine learning in python. *arXiv* 2012 Jan 2. [doi: [10.3389/fninf.2014.00014](https://doi.org/10.3389/fninf.2014.00014)]
36. Pennington J, Socher R, Manning C. Glove: global vectors for word representation. In: *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. 2014 Presented at: 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP); Oct, 2014; Doha, Qatar. [doi: [10.3115/v1/d14-1162](https://doi.org/10.3115/v1/d14-1162)]
37. Joachims T. Text categorization with Support Vector Machines: learning with many relevant features. In: *Machine Learning: ECML-98*. Berlin, Heidelberg: Springer; 1998. [doi: [10.1007/BFb0026683](https://doi.org/10.1007/BFb0026683)]
38. Zhang W, Yoshida T, Tang X. Text classification based on multi-word with support vector machine. *Knowl Based Syst* 2008 Dec;21(8):879-886. [doi: [10.1016/j.knosys.2008.03.044](https://doi.org/10.1016/j.knosys.2008.03.044)]
39. Luss R, D'Aspremont A. Predicting abnormal returns from news using text classification. *Quant Finance* 2012 Mar 29;15(6):999-1012. [doi: [10.1080/14697688.2012.672762](https://doi.org/10.1080/14697688.2012.672762)]
40. Jardino M. Multilingual stochastic n-gram class language models. In: *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing Conference Proceedings*. 1996 Presented at: IEEE International Conference on Acoustics, Speech, and Signal Processing Conference Proceedings; May 9, 1996; Atlanta, GA, USA. [doi: [10.1109/icassp.1996.540315](https://doi.org/10.1109/icassp.1996.540315)]
41. Vig J, Belinkov Y. Analyzing the structure of attention in a transformer language model. In: *Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*. 2019 Presented at: 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP; Aug 1, 2019; Florence, Italy. [doi: [10.18653/v1/w19-4808](https://doi.org/10.18653/v1/w19-4808)]
42. Shen J, Rudzicz F. Detecting anxiety through reddit. In: *Proceedings of the Fourth Workshop on Computational Linguistics and Clinical Psychology — From Linguistic Signal to Clinical Reality*. 2017 Presented at: Fourth Workshop on Computational Linguistics and Clinical Psychology — From Linguistic Signal to Clinical Reality; Aug 3, 2017; Vancouver, BC. [doi: [10.18653/v1/w17-3107](https://doi.org/10.18653/v1/w17-3107)]
43. Burdisso SG, Errecalde M, Montes-y-Gómez M. Using text classification to estimate the depression level of reddit users. *J Comput Sci Technol* 2021 Apr 17;21(1):e1. [doi: [10.24215/16666038.21.e1](https://doi.org/10.24215/16666038.21.e1)]
44. Devlin J, Chang M, Lee K, Toutanova K. BERT: pre-training of deep bidirectional transformers for language understanding. In: *Proceedings of NAACL-HLT 2019*. 2019 Presented at: NAACL-HLT 2019; Jun 2 - 7, 2019; Minneapolis, Minnesota URL: <https://arxiv.org/abs/1810.04805>
45. Lan Z, Chen M, Goodman S, Gimpel K, Sharma P, Soricut R. Albert: a lite BERT for self-supervised learning of language representations. In: *Proceedings of the 8th International Conference on Learning Representations, ICLR 2020*. 2020 Presented at: 8th International Conference on Learning Representations, ICLR 2020; Apr 26-30, 2020; Addis Ababa, Ethiopia URL: <https://arxiv.org/abs/1909.11942>
46. Lee J, Yoon W, Kim S, Kim D, Kim S, So CH, et al. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics* 2020 Feb 15;36(4):1234-1240 [FREE Full text] [doi: [10.1093/bioinformatics/btz682](https://doi.org/10.1093/bioinformatics/btz682)] [Medline: [31501885](https://pubmed.ncbi.nlm.nih.gov/31501885/)]
47. Beltagy I, Peters M, Cohan A. Longformer: the long-document transformer. *arXiv* 2020 [FREE Full text]
48. Ji S, Zhang T, Ansari L, Fu J, Tiwari P, Cambria E. Mentalbert: publicly available pretrained language models for mental healthcare. In: *Proceedings of the Thirteenth Language Resources and Evaluation Conference*. 2022 Presented at: Thirteenth Language Resources and Evaluation Conference; Jun 20-25, 2022; Marseille, France. [doi: [10.1016/b978-0-323-90118-5.00006-0](https://doi.org/10.1016/b978-0-323-90118-5.00006-0)]
49. Yu X, Hu W, Lu S, Sun X, Yuan Z. BioBERT based named entity recognition in electronic medical record. In: *Proceedings of the 10th International Conference on Information Technology in Medicine and Education (ITME)*. 2019 Presented at: 10th International Conference on Information Technology in Medicine and Education (ITME); Aug 23-25, 2019; Qingdao, China. [doi: [10.1109/itme.2019.00022](https://doi.org/10.1109/itme.2019.00022)]
50. Alghanmi I, Espinosa-Anke L, Schockaert S. Interpreting patient descriptions using distantly supervised similar case retrieval. In: *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 2022 Presented at: SIGIR '22: The 45th International ACM SIGIR Conference on Research and Development in Information Retrieval; Jul 11 - 15, 2022; Madrid Spain. [doi: [10.1145/3477495.3532003](https://doi.org/10.1145/3477495.3532003)]
51. Bai Y, Zhou X. Automatic detecting for health-related twitter data with biobert. In: *Proceedings of the Fifth Social Media Mining for Health Applications Workshop & Shared Task*. 2020 Presented at: Fifth Social Media Mining for Health Applications Workshop & Shared Task; Online; Barcelona, Spain URL: <https://aclanthology.org/2020.smm4h-1.10>

52. González-Carvajal S, Garrido-Merchán E. Comparing BERT against traditional machine learning text classification. arXiv 2021 [[FREE Full text](#)]
53. Clavié B, Alphonsus M. The unreasonable effectiveness of the baseline: discussing SVMs in legal text classification. In: Volume 346: Legal Knowledge and Information Systems. Amsterdam: IOS Press; 2021. URL: <https://tinyurl.com/4dtkv9rt>
54. Simple transformers homepage. Simple Transformers. URL: <https://simpletransformers.ai/> [accessed 2021-01-04]
55. Wolf T, Debut L, Sanh V, Chaumond J, Delangue C, Moi A, et al. Transformers: state-of-the-art natural language processing. In: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations. 2020 Presented at: 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations; Oct, 2020; Online. [doi: [10.18653/v1/2020.emnlp-demos.6](https://doi.org/10.18653/v1/2020.emnlp-demos.6)]
56. Classification specifics. Simple Transformers. URL: <https://simpletransformers.ai/docs/classification-specifics/#dealing-with-long-text> [accessed 2021-04-15]
57. Wang X, Zhang C, Ji Y, Sun L, Wu L, Bao Z. A depression detection model based on sentiment analysis in micro-blog social network. In: Trends and Applications in Knowledge Discovery and Data Mining. Berlin, Heidelberg: Springer; 2013. [doi: [10.1007/978-3-642-40319-4_18](https://doi.org/10.1007/978-3-642-40319-4_18)]
58. Hassan A, Hussain J, Hussain M, Sadiq M, Lee S. Sentiment analysis of social networking sites (SNS) data using machine learning approach for the measurement of depression. In: Proceedings of the 2017 International Conference on Information and Communication Technology Convergence (ICTC). 2017 Presented at: 2017 International Conference on Information and Communication Technology Convergence (ICTC); Oct 18-20, 2017; Jeju, Korea (South). [doi: [10.1109/ictc.2017.8190959](https://doi.org/10.1109/ictc.2017.8190959)]
59. Stephen JJ, Prabhu P. Detecting the magnitude of depression in Twitter users using sentiment analysis. Int J Electrical Comput Eng 2019 Aug 01;9(4):3247. [doi: [10.11591/ijece.v9i4.pp3247-3255](https://doi.org/10.11591/ijece.v9i4.pp3247-3255)]
60. Liu T, Meyerhoff J, Eichstaedt JC, Karr CJ, Kaiser SM, Kording KP, et al. The relationship between text message sentiment and self-reported depression. J Affect Disord 2022 Apr 01;302:7-14. [doi: [10.1016/j.jad.2021.12.048](https://doi.org/10.1016/j.jad.2021.12.048)] [Medline: [34963643](https://pubmed.ncbi.nlm.nih.gov/34963643/)]
61. Pota M, Ventura M, Catelli R, Esposito M. An effective BERT-based pipeline for Twitter sentiment analysis: a case study in Italian. Sensors (Basel) 2020 Dec 28;21(1):133 [[FREE Full text](#)] [doi: [10.3390/s21010133](https://doi.org/10.3390/s21010133)] [Medline: [33379231](https://pubmed.ncbi.nlm.nih.gov/33379231/)]
62. Hutto C, Gilbert E. VADER: a parsimonious rule-based model for sentiment analysis of social media text. Proc Int AAI Conf Web Social Media 2014 May 16;8(1):216-225. [doi: [10.1609/icwsm.v8i1.14550](https://doi.org/10.1609/icwsm.v8i1.14550)]
63. Nguyen D, Vu T, Nguyen A. BERTweet: a pre-trained language model for English Tweets. In: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations. 2020 Presented at: Conference on Empirical Methods in Natural Language Processing: System Demonstrations; Nov 16-20, 2020; Online. [doi: [10.18653/v1/2020.emnlp-demos.2](https://doi.org/10.18653/v1/2020.emnlp-demos.2)]
64. Bates D, Mächler M, Bolker B, Walker S. Fitting linear mixed-effects models using lme4. J Stat Softw 2015;67(1):1-48. [doi: [10.18637/jss.v067.i01](https://doi.org/10.18637/jss.v067.i01)]
65. Wood SN, Pya N, Säfken B. Smoothing parameter and model selection for general smooth models. J Am Statistical Assoc 2017 Jan 04;111(516):1548-1563. [doi: [10.1080/01621459.2016.1180986](https://doi.org/10.1080/01621459.2016.1180986)]
66. The R project for statistical computing. CRAN R Project. URL: <https://www.r-project.org/> [accessed 2022-07-13]
67. Matloff N. Statistical Regression and Classification From Linear Models to Machine Learning. Boca Raton, Florida, United States: CRC Press; 2017. URL: <https://tinyurl.com/4sfjbp9t>
68. van der Brug W. Issue ownership and party choice. Electoral Stud 2004 Jun;23(2):209-233. [doi: [10.1016/s0261-3794\(02\)00061-6](https://doi.org/10.1016/s0261-3794(02)00061-6)]
69. Guo H, Zhi W, Liu H, Xu M. Imbalanced learning based on logistic discrimination. Comput Intell Neurosci 2016;2016:5423204 [[FREE Full text](#)] [doi: [10.1155/2016/5423204](https://doi.org/10.1155/2016/5423204)] [Medline: [26880877](https://pubmed.ncbi.nlm.nih.gov/26880877/)]
70. Sokolova M, Japkowicz N, Szpakowicz S. Beyond accuracy, F-Score and ROC: a family of discriminant measures for performance evaluation. In: AI 2006: Advances in Artificial Intelligence. Berlin, Heidelberg: Springer; 2006. [doi: [10.1007/11941439_114](https://doi.org/10.1007/11941439_114)]
71. Blei D, Ng A, Jordan M. Latent dirichlet allocation. J Mach Learn Res 2003 Mar 1;3:993-1022 [[FREE Full text](#)]
72. Mallet: machine learning for language toolkit homepage. Mallet: MACHine Learning for Language Toolkit. URL: <http://mallet.cs.umass.edu> [accessed 2022-04-19]
73. Howes C, Purver M, McCabe R. Linguistic indicators of severity and progress in online text-based therapy for depression. In: Proceedings of the Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality. 2014 Presented at: Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality; Jun, 2014; Baltimore, Maryland, USA. [doi: [10.3115/v1/w14-3202](https://doi.org/10.3115/v1/w14-3202)]
74. de Choudhury M, Gamon M, Counts S, Horvitz E. Predicting depression via social media. Proc Int AAI Conference Web Social Media 2021 Aug 03;7(1):128-137. [doi: [10.1609/icwsm.v7i1.14432](https://doi.org/10.1609/icwsm.v7i1.14432)]
75. van de Leemput IA, Wichers M, Cramer AO, Borsboom D, Tuerlinckx F, Kuppens P, et al. Critical slowing down as early warning for the onset and termination of depression. Proc Natl Acad Sci U S A 2014 Jan 07;111(1):87-92 [[FREE Full text](#)] [doi: [10.1073/pnas.1312114110](https://doi.org/10.1073/pnas.1312114110)] [Medline: [24324144](https://pubmed.ncbi.nlm.nih.gov/24324144/)]
76. Couto M, Pérez A, Parapar J. Temporal word embeddings for early detection of signs of depression. In: Proceedings of the CIRCLE (Joint Conference of The Information Retrieval Communities in Europe). 2022 Presented at: CIRCLE (Joint

- Conference of The Information Retrieval Communities in Europe); Jul 04-07 2022; Toulouse, Fr URL: https://ceur-ws.org/Vol-3178/CIRCLE_2022_paper_03.pdf
77. Tausczik YR, Pennebaker JW. The psychological meaning of words: LIWC and computerized text analysis methods. *J Lang Social Psychol* 2009 Dec 08;29(1):24-54. [doi: [10.1177/0261927x09351676](https://doi.org/10.1177/0261927x09351676)]
78. Reddit Self-reported Depression Diagnosis (RSDD) dataset. Georgetown University. URL: <https://ir.cs.georgetown.edu/resources/rsdd.html> [accessed 2023-02-28]

Abbreviations

ALBERT: A Lite Bidirectional Encoder Representations from Transformers

BERT: Bidirectional Encoder Representations from Transformers

LM: language model

MDD: major depressive disorder

NLP: natural language processing

RSDD: Reddit Self-reported Depression Diagnosis

SVM: support vector machine

TF-IDF: term frequency–inverse document frequency

Edited by K El Emam; submitted 23.07.22; peer-reviewed by A Teles, T Zhang; comments to author 11.11.22; revised version received 06.01.23; accepted 15.01.23; published 24.03.23.

Please cite as:

Owen D, Antypas D, Hassoulas A, Pardiñas AF, Espinosa-Anke L, Collados JC

Enabling Early Health Care Intervention by Detecting Depression in Users of Web-Based Forums using Language Models: Longitudinal Analysis and Evaluation

JMIR AI 2023;2:e41205

URL: <https://ai.jmir.org/2023/1/e41205>

doi: [10.2196/41205](https://doi.org/10.2196/41205)

PMID: [37525646](https://pubmed.ncbi.nlm.nih.gov/37525646/)

©David Owen, Dimosthenis Antypas, Athanasios Hassoulas, Antonio F Pardiñas, Luis Espinosa-Anke, Jose Camacho Collados. Originally published in JMIR AI (<https://ai.jmir.org>), 24.03.2023. This is an open-access article distributed under the terms of the Creative Commons Attribution License (<https://creativecommons.org/licenses/by/4.0/>), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR AI, is properly cited. The complete bibliographic information, a link to the original publication on <https://www.ai.jmir.org/>, as well as this copyright and license information must be included.

Original Paper

Machine Learning for the Prediction of Procedural Case Durations Developed Using a Large Multicenter Database: Algorithm Development and Validation Study

Samir Kendale^{1*}, MD; Andrew Bishara^{2,3*}, MD; Michael Burns^{4*}, MD, PhD; Stuart Solomon⁵, MD; Matthew Corriere^{6*}, MD; Michael Mathis^{4,7*}, MD

¹Department of Anesthesia, Critical Care & Pain Medicine, Beth Israel Deaconess Medical Center, Boston, MA, United States

²Department of Anesthesia and Perioperative Care, University of California, San Francisco, San Francisco, CA, United States

³Bakar Computational Health Sciences Institute, University of California San Francisco, San Francisco, CA, United States

⁴Department of Anesthesiology, University of Michigan Medical School, Ann Arbor, MI, United States

⁵Department of Anesthesiology, The University of Texas Health Science Center at San Antonio, San Antonio, TX, United States

⁶Department of Surgery, Section of Vascular Surgery, University of Michigan Medical School, Ann Arbor, MI, United States

⁷Center for Computational Medicine and Bioinformatics, University of Michigan Medical School, Ann Arbor, MI, United States

*these authors contributed equally

Corresponding Author:

Samir Kendale, MD

Department of Anesthesia, Critical Care & Pain Medicine

Beth Israel Deaconess Medical Center

1 Deaconess Road

Boston, MA, 02215

United States

Phone: 1 6177545400

Email: skendale@bidmc.harvard.edu

Abstract

Background: Accurate projections of procedural case durations are complex but critical to the planning of perioperative staffing, operating room resources, and patient communication. Nonlinear prediction models using machine learning methods may provide opportunities for hospitals to improve upon current estimates of procedure duration.

Objective: The aim of this study was to determine whether a machine learning algorithm scalable across multiple centers could make estimations of case duration within a tolerance limit because there are substantial resources required for operating room functioning that relate to case duration.

Methods: Deep learning, gradient boosting, and ensemble machine learning models were generated using perioperative data available at 3 distinct time points: the time of scheduling, the time of patient arrival to the operating or procedure room (primary model), and the time of surgical incision or procedure start. The primary outcome was procedure duration, defined by the time between the arrival and the departure of the patient from the procedure room. Model performance was assessed by mean absolute error (MAE), the proportion of predictions falling within 20% of the actual duration, and other standard metrics. Performance was compared with a baseline method of historical means within a linear regression model. Model features driving predictions were assessed using Shapley additive explanations values and permutation feature importance.

Results: A total of 1,177,893 procedures from 13 academic and private hospitals between 2016 and 2019 were used. Across all procedures, the median procedure duration was 94 (IQR 50-167) minutes. In estimating the procedure duration, the gradient boosting machine was the best-performing model, demonstrating an MAE of 34 (SD 47) minutes, with 46% of the predictions falling within 20% of the actual duration in the test data set. This represented a statistically and clinically significant improvement in predictions compared with a baseline linear regression model (MAE 43 min; $P < .001$; 39% of the predictions falling within 20% of the actual duration). The most important features in model training were historical procedure duration by surgeon, the word “free” within the procedure text, and the time of day.

Conclusions: Nonlinear models using machine learning techniques may be used to generate high-performing, automatable, explainable, and scalable prediction models for procedure duration.

KEYWORDS

medical informatics; artificial intelligence; AI; machine learning; operating room; OR management; perioperative; algorithm development; validation; patient communication; surgical procedure; prediction model

Introduction

Background

Across health care settings, anesthesiologist staffing and resources are commonly allocated based on procedure volume, concurrency, complexity, and projected duration [1,2]. Although preparing allocations is often done far in advance, depending on institutional processes, daily scheduling requires accurate information regarding recovery room availability as well as surgical, anesthesiology, and nurse staffing, all of which directly rely on accurate determination of procedure duration. More accurate prediction of procedure duration may allow for more effective assignment of procedure rooms, more efficient scheduling of cases (eg, staggering procedure rooms for surgeons with multiple cases), more predictable hours for involved staff, and clearer patient communication. Firmer understanding also relates to the high cost of running procedure rooms and maintaining optimal procedure room use. In addition, inaccurate estimates of case length affect patient care because they lead to gaps within block schedules that are not optimally used. This can lead to add-on cases not being completed in a timely manner as well as bed control issues in the inpatient setting or discharge issues in the outpatient setting. To manage procedure time, most institutions use either surgeon-directed procedure durations or procedure durations based on historical averages [3,4], which can be frequently inaccurate [1,2]. Because of the complexity of the problem and the inclusion of large numbers of features with potential interactions, linear regression methods to predict procedure durations have demonstrated varying levels of success [5-9]. Machine learning approaches have been proposed to mitigate this issue. In short, machine learning aims to extract patterns of knowledge from data, the benefit being the ability to process large volumes of disparate data, exploring potentially nonlinear interactions that may challenge the required assumptions of conventional analysis. Nonetheless, current studies have been limited to single or few institutions, smaller sample sizes (between 400 and 80,000 cases), specific surgical subpopulations (robotic [10], colorectal [11], and pediatric [12]), or the use of proprietary algorithms [10-16]; for example, the study by Lam et al [11] was multicenter but had approximately 10,000 colorectal cases. The studies by Tuwatananurak et al [13] and Rozario and Rozario [14] used proprietary tools, which may be useful for adoption but do not permit the same level of transparency or explainability as other methods. The included features varied significantly across previous studies.

Objectives

Given the limitations of previous studies and the dependency of machine learning performance on training set size and heterogeneity, we developed a machine learning algorithm derived from a large multicenter data set for a more accurate prediction of surgical procedure duration compared with

historical averages of procedure time. We hypothesized that a machine learning algorithm derived from a large multicenter data set with >1 million procedures would more accurately predict surgical procedure duration than a baseline linear regression approach. Using an explainable machine learning-based algorithm, the results can provide additional valuable insight regarding procedure duration and variability. The clinical objective of this protocol was to determine whether a machine learning algorithm scalable across multiple centers could make estimations of case duration within a tolerance limit because there are substantial resources required for procedure room functioning that relate to case duration.

Methods

Ethics Approval

We obtained institutional review board approval for this multicenter observational study from New York University (NYU) Langone Health, New York, NY (S19-01451), and the requirement for written informed consent was waived by the institutional review board.

Study Design

We followed multidisciplinary guidelines for reporting machine learning-based prediction models in biomedical research [17,18]. Study outcomes, data collection, and statistical methods were established a priori and presented and approved at a multicenter peer review forum on January 13, 2020, before data analyses [19].

Data Source

Data were provided by the Multicenter Perioperative Outcomes Group (MPOG). Within this research consortium, data from enterprise and departmental electronic health record systems are routinely uploaded to a secure centralized database. Methods for local electronic health record data acquisition, validation, mapping to interoperable universal MPOG concepts, and secure transfer to the coordinating center have been previously described [20] and used in multiple published studies [21-24]. In brief, each center uses a standardized set of data diagnostics to evaluate and address data quality on a monthly basis. Random subsets of cases are manually audited by a clinician at each center to assess, and attest to, the accuracy of data extraction and source data. At each institution participating in the MPOG, at the time of clinical onboarding (ie, when a new site joins the MPOG), a site-level data audit that involves hundreds of cases is initially performed until reaching a level of accuracy acceptable to the local site data quality reviewer. After this iterative process, the onboarded sites undergo a manual review of a minimum of 5 cases per month to ensure that changes in clinical and documentation practice patterns do not meaningfully degrade data quality over time [20]. All institutions were in the United States and ranged from community hospitals to large

academic centers. A list of included centers is provided in Table S1 of [Multimedia Appendix 1](#).

Study Population

The study population included adult and pediatric patients who underwent procedures requiring anesthesiologist care between June 1, 2016, and November 30, 2019. Labor epidurals, labor analgesia, and procedures lacking relevant time points (patient-in-room duration) or provider information (surgeon and anesthesia staff identities anonymized) were excluded. Other missing data were handled as described in the following subsections.

Primary Outcome

The primary outcome was procedure duration. Procedure duration was defined according to the precomputed *procedure room duration* electronic health record phenotype, interoperable across a wide variety of electronic health record vendors. The implications of over- and underpredicting the length of the procedure cannot be universally defined because this will be dictated by institutional policy and culture, but, broadly, overprediction (predicting a longer case than actual duration) may result in underuse of a given surgical block time, whereas

underprediction (predicting a shorter case than actual duration) may result in inadequate staffing models.

Basic Model Features

The features considered were determined by availability within the MPOG data and included certain patient characteristics such as sex, height, weight, and BMI; medical comorbidities; allergies; baseline vital signs; functional status; home medications; the day of the week; procedure text; procedure room type; anesthesia techniques; case times and durations; and deidentified institution and staff identities. Features were selected for modeling based on a review of the existing literature as well as by clinical and managerial experience [6,8,25]. [Table 1](#) indicates the features that were ultimately selected to be used for the primary model and sensitivity analysis models using data available at varying time points relative to the start of the procedure. The primary model used features only available at the time the patient arrived in the procedure room. Of the 2 secondary models, one used features restricted to those available at the time of surgical scheduling, and the other used features expanded to those available after patient arrival to the procedure room up to the time of procedure start. This is described further in the Sensitivity Analyses subsection.

Table 1. Summary of prediction model features.

Model feature	Included in <i>Time of Scheduling</i> model (secondary model)	Included in <i>Time of Patient in OR</i> ^a model (primary model)	Included in <i>Time of Surgical Incision</i> model (secondary model)
Case duration	✓	✓	✓
Holiday	✓	✓	✓
Weekend	✓	✓	✓
Surgical service	✓	✓	✓
Surgical procedure text	✓	✓	✓
Anonymized surgeon identity	✓	✓	✓
Patient age	✓	✓	✓
Patient BMI	✓	✓	✓
Location type (acute care hospital, mixed use OR, freestanding ambulatory surgical center, etc)	✓	✓	✓
Institution	✓	✓	✓
Preoperative comorbidities, including arrhythmia, CHF ^b , CAD ^c , HTN ^d , MI ^e , COPD ^f , diabetes, renal failure, liver disease, coagulopathy, cancer, and psychiatric illness (based on MPOG ^g phenotype or preoperative anesthesia H&P ^h)	✓	✓	✓
Number of allergies	✓	✓	✓
Preoperative laboratory values, including creatinine, hemoglobin, albumin, INR ⁱ , and glucose levels		✓	✓
Preoperative baseline blood pressure		✓	✓
Preoperative existing airway		✓	✓
Anonymized anesthesia staff		✓	✓
ASA ^j physical status score		✓	✓
Type of anesthesia			✓
Type of airway management			✓
Presence of nerve block			✓
Presence of neuraxial block			✓
Number of intravenous lines at the time of surgical procedure start			✓
Presence of arterial line at the time of surgical procedure start			✓
Time from patient arrival in the OR to anesthesia induction end			✓
Time from anesthesia induction end to surgical incision			✓

^aOR: operating room.

^bCHF: congestive heart failure.

^cCAD: coronary artery disease.

^dHTN: hypertension.

^eMI: myocardial infarction.

^fCOPD: chronic obstructive pulmonary disease.

^gMPOG: Multicenter Perioperative Outcomes Group.

^hH&P: history and physical examination.

ⁱINR: international normalized ratio.

^jASA: American Society of Anesthesiologists.

Experience and Historical Features

Additional features were derived from surgical staff and institution identity (Textbox 1).

Derived experience and historical features were computed on a monthly basis from the earliest available date to the month before a given procedure; for example, procedures for the month of February 2019 would include derived features from June 2016 (the first available date in the data set) until January 2019. Procedure-wise features (eg, features derived from the first available date until the actual date of the procedure) were not included in the data set owing to computational processing power cost. Only the primary surgeon identity, anesthesiologist identity, and current procedural terminology (CPT) code were used in feature engineering. Density features were included to

account for surgeons or institutions that were not included from the earliest date in the MPOG data set; for example, a surgeon who performed 20 procedures in 2 months would have the same density as a surgeon who performed 40 procedures in 4 months to mitigate model bias attributable to surgeons or institutions first appearing in the data set beyond the start date of the data set. Surgeon and institution experience are limited by the start date of the data set and would not account for experience before this start date. The same surgeon would have 2 different identities at different facilities because surgeons may fundamentally do different things based on the hospital they are practicing at, given the resources available to them at that specific hospital and practice patterns that are generally followed at that hospital.

Textbox 1. Additional features.

- Surgeon experience: total number of procedures performed by surgeon
- Surgeon procedure experience: total number of a given procedure (by anesthesiology current procedural terminology [CPT] code) performed by surgeon
- Institutional procedure experience: total number of a given procedure performed at an institution
- Historical procedure duration: historical mean duration of a given procedure (by CPT code)
- Historical procedure duration by institution: historical mean duration of a given procedure at an institution
- Historical procedure duration by surgeon: historical mean duration of a given procedure by surgeon
- Surgeon total density: surgeon experience divided by time since surgeon's first procedure
- Surgeon procedure density: surgeon procedure experience divided by time since surgeon's first procedure
- Institutional procedure density: institutional procedure experience divided by time since institution's first procedure

Procedure Text Features

Although it was an option to include only the machine learning-generated anesthesia CPT code as a feature, it was felt that these codes lack the granularity that would be needed for more accurate prediction in this context. *Procedure text* refers to the name of the surgical procedure as booked by the surgeon. As the data set was being generated from a variety of institutions, *procedure text* may refer to either a scheduled procedure or a performed procedure and may vary in descriptiveness based on surgeon preference and institutional culture. Examples of *procedure text* may be “laparoscopic cholecystectomy with intraoperative cholangiogram” or “posterior cervical fusion C3-C7.” Natural language processing was used to convert text into a form usable by machine learning. Through a manual review of the corpus, common misspellings were corrected, and the 5 most common abbreviations were expanded (as detailed in Multimedia Appendix 1 [refer to R Code for Data Processing]). To decrease vocabulary size, text was standardized through the removal of punctuations and common stop words (eg, “a,” “an,” and “the”). Additional words deemed likely to be nondeterminative of procedure duration, such as “right” and “left,” were also preemptively removed. After text processing, term matrices were created with 1- and 2-word n-grams. Term frequency-inverse document frequency was used to transform text into numerical values. Because of the vastness of the corpus, but also to retain as many relevant terms as possible, terms with document frequency >0.995 were

removed because these terms likely did not contain important information. Similar processing of procedure text for machine learning has been described in other published works [21]. The code for natural language processing is provided along with other data processing code in Multimedia Appendix 1 (refer to R Code for Data Processing).

Power Analysis

Previous studies estimating procedure duration have used between 400 and 80,000 cases [9,10,12,13,26]. On the basis of experience and other comparable machine learning problems, we estimated that at least 100,000 cases encompassing a wide range of surgical procedure types would be adequate. On the basis of initial cohort size queries, there were >100,000 cases available for training, testing, and validation. A greater number of cases with a wider diversity in procedure types leads to a stronger machine learning model with less overfitting and ultimately greater generalizability [27,28].

Data Preprocessing

All data were examined for missingness and veracity; cases with missing procedure duration and surgeon or anesthesia staff identities were eliminated. Outlier cases with durations of >1440 minutes were removed. Any feature missing >40% of the values or missing from >40% of the institutions was excluded. The remaining features were considered qualifying data. Different machine learning algorithms automatically treated missing values differently: generalized linear models use mean

imputation to handle missing data. The missing values are replaced with the mean of the nonmissing values for that feature. Gradient boosting machines learn optimal splits in the decision trees for missing values. These algorithms do not impute missing data; instead, they find the best path in the decision trees for the observations with missing values. Deep learning algorithms perform mean imputation by default for handling missing data; they replace the missing values with the mean of the nonmissing values for that feature during training. Machine learning packages used for modeling also reject unimportant features, and this functionality was retained during modeling.

Statistical Analyses: Model Development

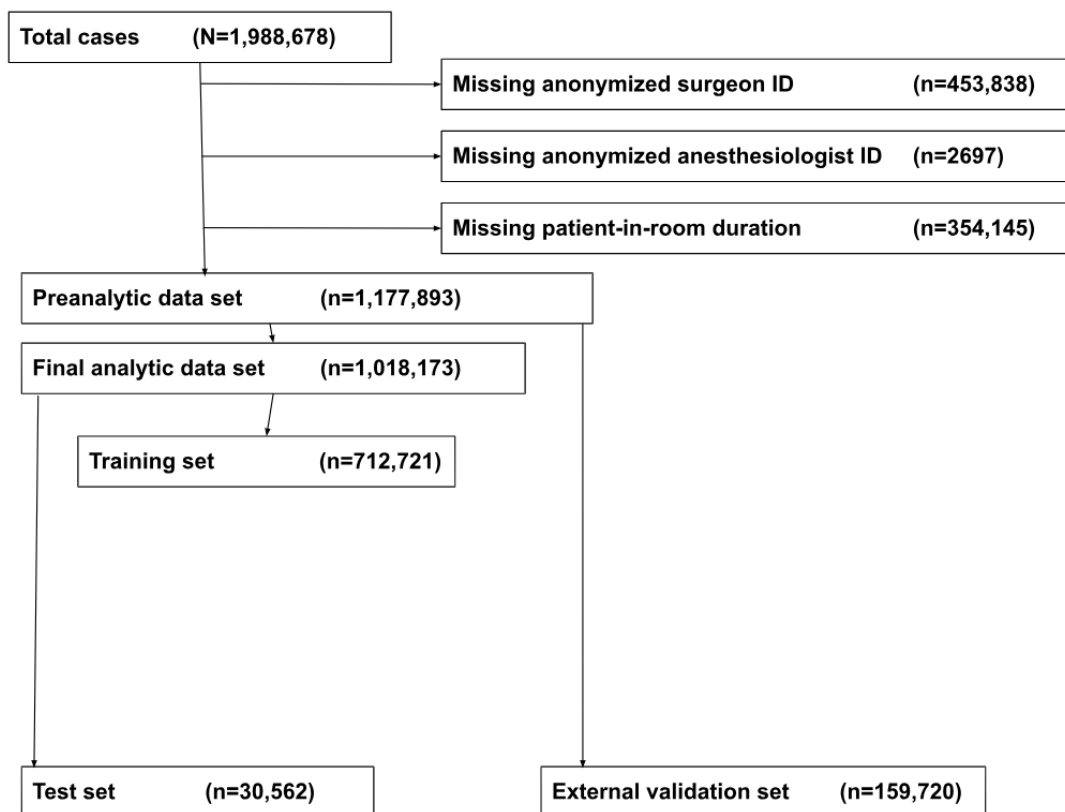
The primary model was designed using a temporal reference point of patient arrival to the procedure room and thus only used data available before this event. The analysis was performed in R statistical software (R Foundation for Statistical Computing), using the *H2O* package of machine learning algorithms [29]. The models were generated and run via a server with a 2.9 GHz Intel processor with 96 GB RAM and a 64-bit operating system. This is comparable to a standard hospital computer system. Categorical variables were automatically processed to one-hot encoding. Predicted anesthesia CPT codes were used to characterize procedure types, using a previously published prediction model [21]. Multiple supervised machine learning regression algorithms were trained, including deep learning, gradient boosting machine, and stacked ensemble methods. In brief, deep learning helps to identify complex patterns, in which layers of nodes receive input and offer output, with successive layers representing more complex combinations of prior simpler

layers [30]. By contrast, gradient boosting machines use weaker learners, specifically decision trees, by iteratively modifying the weights of each observation and progressively combining the trees together to improve the fit of the model [31]. Finally, stacked ensemble methods use combinations of strong learners (ie, deep learning, gradient boosting, and logistic regression) to optimize performance [32,33]. The best-performing model was further tuned, depending on the available hyperparameters for tuning. Hyperparameter tuning was accomplished using grid search, the default within the *H2O* package. Although the gradient boosting machine model was trained and tuned separately (*h2o.gbm* function in the *H2O* package), the deep learning and stacked ensemble models were generated using an automatic machine learning method (*autoML* function in the *H2O* package), which created and compared 10 distinct machine learning models.

Data Partitioning

Split-set validation was used, in which 70% of the data were used for training and 30% for testing. Internal validation was additionally performed by using 5-fold cross-validation on the training set. In k-fold cross-validation, the training data set is divided into k subsets or *folds*. Each fold acts as a validation set for a specific model, whereas the remaining k-1 folds are used to train the model. This process is repeated k times, with each fold being used as a validation set exactly once. The model performance is then averaged over all k iterations. Data from 1 randomly selected institution were not included in the training or test sets and were used as a true *holdout* data set for external validation to further assess model generalizability (Figure 1).

Figure 1. Study inclusion and exclusion criteria and machine learning model training and validation and testing schematic.



Performance Metrics

Model performance was assessed primarily using mean absolute error (MAE) and root mean square error (RMSE). In addition, for comparison with other published models and to further account for both distribution and outliers, median absolute error with IQR and mean absolute percentage error (MAPE) were also calculated. Because of high dimensional distributions and assumptions required for the generation of prediction intervals, two methods of assessing procedure duration variance were used: (1) a second model was trained, using the absolute error of each procedure prediction in the training set as the target output, and this model was then applied to the test set to generate a prediction error for each test set case; and (2) the loss function was modified to a quantile distribution, and 2 additional models were trained at values of 0.2 and 0.8 quantiles. A bootstrap method with 1000 repetitions was also attempted for generating a prediction interval, although we anticipated and confirmed that this was computationally expensive and time consuming beyond the utility of the workflow necessities of this algorithm.

The tuned best-performing final machine learning model was compared with a common historical reference model: historical procedure time by surgeon as the sole feature (independent variable) of a linear regression model. This was the same derived feature included in the machine learning models [5,34]. This feature was selected for comparison because *historical procedure time by surgeon* is commonly used by many institutions as the sole variable in their prediction models when cases are booked into the procedure room schedule. A comparison was performed using the Wilcoxon rank sum test on the model errors. Other available approaches such as Bayesian methods were not used for comparison owing to differences in the intended implementation, the availability of certain factors such as surgeon-estimated operative times, and the requirement to significantly modify the data structure.

All models were assessed for the distribution of error, overage (how frequently actual duration exceeded predicted duration), underage (how frequently actual duration underestimated predicted duration), and the percentage of procedures in which the predicted duration fell within 20% of the actual duration. Overage and underage are useful for broadly understanding whether the models tend to overestimate or underestimate the prediction. For each generated prediction interval (either predicted error or quantile loss function models), the percentage of procedures within the predicted range is also included as a performance metric. As performance metrics for procedure duration calculation vary widely in the literature and are often challenging to interpret by practicing clinicians and procedure room managers, we surveyed several procedure room managers to determine the most intuitive and useful metrics for use in a real-world setting. Finally, model use times were assessed to confirm that high-performing models are not too computationally intensive for practical use.

Model Explainability Subanalysis

To facilitate improved explainability for applicable models, global and local plots of Shapley additive explanations (SHAP) values were developed [35]. SHAP is a framework built on game theory that provides greater interpretability of machine

learning models. Global visualizations included permutation feature importance and SHAP global summary dot plots [36]. SHAP global summary dot plots relate the value of features to the outcome, as opposed to permutation feature importance, which relates the value of the feature to a selected performance metric. The SHAP value indicates how the value of a feature for a given procedure contributed to the prediction. A positive SHAP value contribution indicates that a feature increased the prediction above the average value, whereas a negative SHAP value contribution indicates that a feature decreased the prediction below the average value.

In addition, sample outputs were developed, including predicted duration, prediction interval, and SHAP local plots indicating the features, including direction and magnitude, that affected the output most for a given procedure. Similar approaches for explainability have been used in other medical machine learning applications within health care [37,38].

Sensitivity Analyses

To better characterize the trade-offs between prediction model actionability and accuracy, 2 additional models were generated for use at different time points. One model used features restricted to those available at the time of surgical scheduling, and the other model used features expanded to those available after patient arrival to the procedure room, up to the time of procedure start (eg, surgical incision for operative procedures). Table 1 describes the models that were developed and the features that were determined available for use in the models.

To characterize the extent to which longer procedures influenced the results, 2 secondary subgroup analyses were performed, restricted to procedures lasting <180 minutes and <120 minutes. These 2 subgroup analyses were selected as clinically practical choices from the perspective of procedure room scheduling administrators. Given that longer procedures would likely be associated with greater error in prediction, this would provide an indication of the performance of shorter procedures.

Results

Population Baseline Characteristics

The training and testing data set included 1,018,173 unique procedures across 13 institutions, and the holdout data set included 159,720 procedures from a single institution (Figure 1). The number of cases at each deidentified center is provided in Table S2 in Multimedia Appendix 1. The median procedure duration was 94 (IQR 50-167) minutes; the 5th and 95th percentile durations were 21 and 361 minutes, respectively. Study population baseline characteristics, summarizing all features included in the models, are available in Table S3 in Multimedia Appendix 1. Creatinine, albumin, and international normalized ratio levels exceeded the 40% missing data threshold and were not included in further analyses.

Primary Model Performance Metrics

After modeling and hyperparameter tuning, both the stacked ensemble model and the gradient boosting machine model resulted in comparable performance, with MAEs of 33 minutes and 34 minutes, respectively. The deep learning model

demonstrated an MAE of 35 minutes and an RMSE of 57 minutes at the time of patient arrival to the procedure room, an MAE of 69 minutes and an RMSE of 85 minutes at the time of scheduling, and an MAE of 38 minutes and an RMSE of 62 minutes at the time of incision. The gradient boosting model was selected as the final model because the other performance metrics were comparable, and the tree-based nature of the algorithm allowed for global and local explainability. The final hyperparameters were 500 trees, maximum depth of 5, learning rate of 0.1, stopping tolerance of 0.01, and stopping metric of MAE, with all other hyperparameters at the default setting. The MAE was 19 (IQR 7.5-43) minutes, and the MAPE was 34%. The final model was applied to the single holdout institution for external validation, and model performance metrics are described in Table 2, including an MAE of 38 minutes, which is comparable with the MAE of the test set. For comparison, the performance metrics and specifications of the stacked ensemble model are provided in Tables S4 and S5, respectively, in Multimedia Appendix 1. The linear regression method using

historical procedure time as the sole feature (independent variable) demonstrated an MAE of 43 minutes on the test set and an MAE of 48 minutes on the external validation set and an MAPE of 45%. The difference in error between the linear regression model and the final machine learning model was statistically significant ($P<.001$).

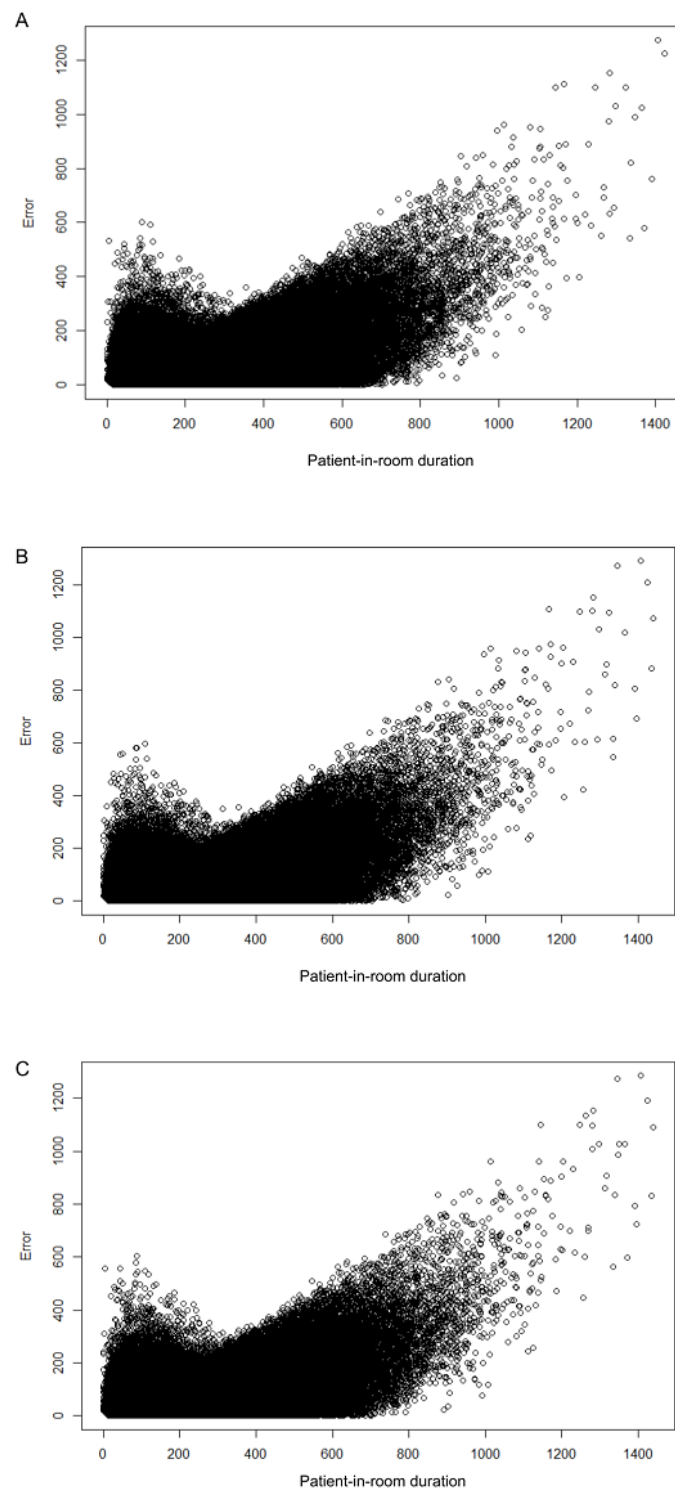
Using 2 different methods for generating prediction intervals, it was determined that the error prediction model resulted in actual procedure times within the predicted range 64% of the time within the primary analysis (Table 2). As anticipated, the bootstrap method was highly computationally expensive (at least 15 min to compute a single prediction interval) and considered impractical for the workflow setting. The prediction intervals of longer-duration procedures were wider than shorter-duration procedures. From observation of the error distribution plots (Figure 2), it seemed clear that longer procedures typically tended to have greater error than shorter procedures. The computation time to predict on the test set (>300,000 cases) was 10 seconds.

Table 2. Performance of optimized surgical duration prediction models at each time point: test set, external validation set, and prediction intervals.

	<i>Time of Scheduling</i> model (secondary model)	<i>Time of Patient in OR^a</i> model (primary model)	<i>Time of Surgical Incision</i> model (secondary model)
Test set			
Mean absolute error (min), mean (SD)	34 (47)	34 (47)	34 (47)
Root mean square error, min	59	59	59
Overage, %	58	58	58
Underage, %	42	42	42
Prediction within 20% of actual duration, %	46	46	46
External validation set			
Mean absolute error (min), mean (SD)	38 (52)	38 (52)	38 (52)
Prediction intervals			
Actual duration within prediction interval, %	Error prediction model: 64; quantile loss function: 61	Error prediction model: 63; quantile loss function: 61	Error prediction model: 65; quantile loss function: 61

^aOR: operating room.

Figure 2. Patient-in-room duration plotted against prediction error. (A) Time of Patient in OR [Operating Room] model (primary model). (B) Time of Scheduling model (secondary model). (C) Time of Surgical Incision model (secondary model).

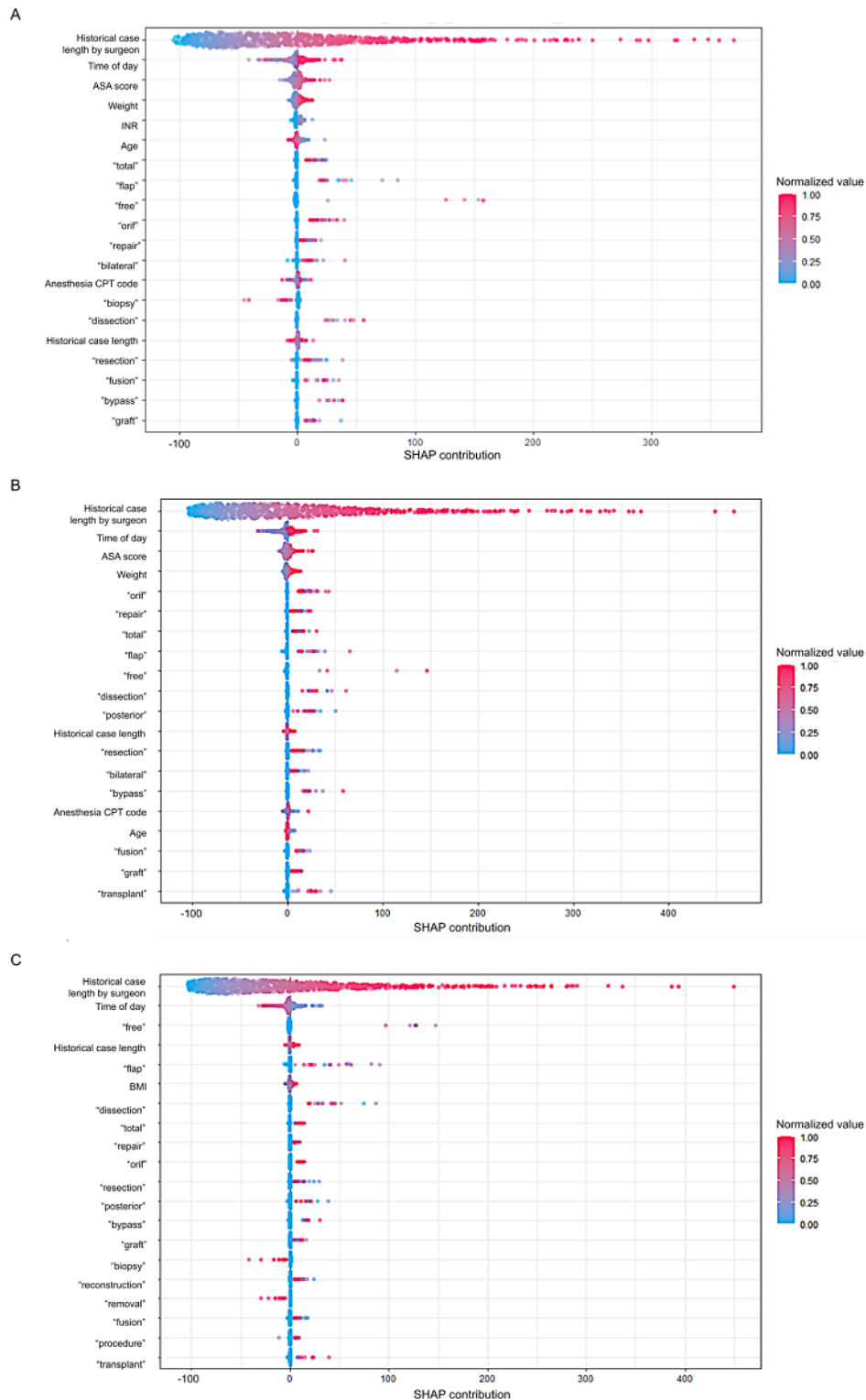


SHAP Global Summary and Feature Importance

The features with the highest importance by feature importance were historical procedure duration by surgeon, the word “free” in the procedure text (eg, “free flap”), and the time of day. The features with the highest importance based on global SHAP

values were historical procedure duration by surgeon, the time of day, and American Society of Anesthesiologists physical status score. SHAP global summary dot plots of each time point model are shown in Figure 3. Permutation feature importance for each time point model is shown in Figure S1 in Multimedia Appendix 1.

Figure 3. Shapley additive explanations (SHAP) global summary dot plots. (A) Time of Patient in OR [Operating Room] model (primary model). (B) Time of Scheduling model (secondary model). (C) Time of Surgical Incision model (secondary model). The feature ranking (y-axis) implies the order of importance of the feature. The SHAP value (x-axis) is a unified index reflecting the impact of a feature on the model output. In each feature importance row, the attributions of all cases to the outcome were plotted using different colored dots, of which the redder dots represent a higher (or positive, if binary) value, and the bluer dots represent a low (or negative, if binary) value, along a gradient from red to blue. ASA: American Society of Anesthesiologists; CPT: current procedural terminology; INR: international normalized ratio.

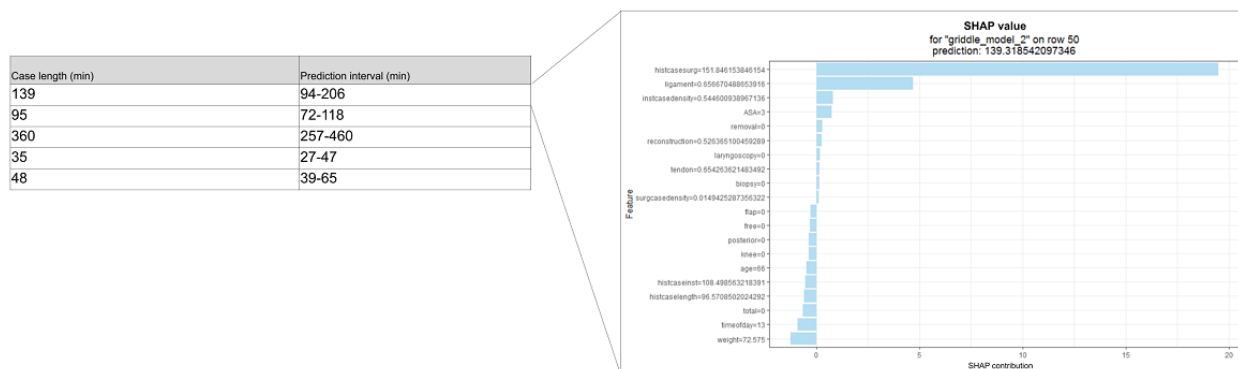


Sample Output

Model outputs feasible for use in real time included predicted time in minutes, the prediction interval as a range, and the SHAP

local explainability plot. As examples, outputs of 5 randomly selected procedures from the test set are shown in Figure 4. For further explanation, a local explainability plot can be easily generated as shown.

Figure 4. Sample output, including Shapley additive explanations (SHAP) local plot. A positive SHAP value contribution indicates that a feature increased the prediction above the average value, whereas a negative SHAP value contribution indicates that a feature decreased the prediction below the average value.



Sensitivity Analyses

In a sensitivity analysis restricted to features available at the time of procedure scheduling, the final model was the gradient boosting machine with an MAE of 34 minutes; for the analysis expanding features to those available up to the time of procedure start (eg, surgical incision), the final model was again the gradient boosting machine with an MAE of 34 minutes (Table 2). At each time point, unused or unimportant columns were dropped by the machine learning algorithm (Table S6 in Multimedia Appendix 1). In the secondary subgroup analyses restricted to shorter procedures, when applying the primary model to procedures lasting <180 minutes, the MAE was 24 minutes, and for procedures lasting <120 minutes, the MAE was 22 minutes.

Discussion

Principal Findings

In this study, we generated machine learning models for the prediction of procedure duration. The final model was the gradient boosting machine, with an MAE of 34 minutes in the test set and an MAE of 38 minutes in the external validation set. This multicenter data set provided a high procedure volume and a wide breadth of procedure types across multiple institutions. Model output included a prediction interval and local explainability for each prediction.

The features with the highest permutation importance were historical procedure duration by surgeon, the word “free” within the procedure text, and the time of day, and those with the highest SHAP values were historical procedure duration by surgeon, American Society of Anesthesiologists physical status score, and the time of day. We speculate that the word “free” having high permutation importance is related to the nature of “free flap” surgery, historically a lengthy procedure. The nonlinear interactions among procedure, surgeon, patient illness severity, and resource availability (the time of day) describe the largest component of the prediction of our model.

Prediction plots suggest that error increases with procedure duration. This result is corroborated by the sensitivity analysis that examined only procedures lasting <120 minutes and <180 minutes, both of which resulted in a lower MAE. Future work might explore different models for different ranges of booking

duration because the models might identify factors in longer procedures that are different from those in shorter procedures.

Supporting Literature

This study expands beyond previous work on single surgical specialties or single-center studies [6,10,13]. Our results show strong performance similarly improved on historical prediction methods [26]. Although it is difficult to compare across different data sets, our model performed grossly better than a single-center model for the prediction of robotic surgery duration (RMSE 80.2 min) [10] and a prospectively evaluated single-center model (MAE 49.5 min) [39]. Compared with a proprietary model tested on 1000 procedures at a single center, which demonstrated a median absolute error of 20 (IQR 10-28) minutes, our model performed comparably with a median absolute error of 19 (IQR 7.5-43) minutes [13]. When compared with a single-center model with a similar analytic approach, our model fared slightly poorer, with a MAPE of 34%, compared with the best surgeon-specific model, with an MAPE of 27% [26]. The approach included multiple surgeon-specific models [26] (as opposed to our unified model, which included all surgeons); considering the high importance of historical case time by surgeon within our model, this difference in performance is expected.

Study Strengths

There have been successful attempts at predicting procedure length, although implementation is often limited by a number of factors, including moderate performance, cumbersome workflow, or the high frequency of unavailable variables. Our major strength is the vast amount of multicenter data. The inherent heterogeneity of practice environments permits potential broader generalizability and customizability of the model, as evidenced by the performance on the test sets and external validation sets. In addition, our approach used commonly available data within the electronic health record that does not rely on human input (ie, human-estimated procedure times), permitting potential improved external implementation.

Our study also introduces several derived features that can be used in other similar projects because the explainability analyses suggest that *historical case length by surgeon*, *institution case density*, *historical case length by institution*, and *historical case length* all have an impact on performance. These features are

relatively simple to compute using the code provided in [Multimedia Appendix 1](#) (refer to Supplementary Code subsection).

Our aim was to develop an easy-to-implement solution with an easy-to-interpret and valuable output. In addition to procedure duration prediction, prediction interval as a measure of variance is useful. Procedures with a high variance can be viewed as less predictable and scheduled accordingly, either after more predictable duration procedures or with no procedures to follow. Furthermore, including an explainability aims to minimize *black box* modeling, building algorithm trust and allowing valuable insight. The global SHAP summary plots improve upon variable importance by relating features to outcome as opposed to relating features to performance metric. The local SHAP plots offer explanations of the drivers of an individual prediction. Although most of these features may not be modifiable, this provides users a data-driven understanding of the drivers of procedure duration [39]; for example, case durations may be *overread* by a procedure room clinician administrator (similar to an electrocardiogram being overread by a cardiologist), and they may be better able to trust or not trust a predicted procedure duration, based on what is most influencing the prediction, and make modifications to procedure room schedule and staffing accordingly. Ideally, the algorithms (similar to most health care artificial intelligence [AI] applications) are used in conjunction with expert opinion and not typically as a sole arbiter of decision-making. In addition, actual cases may on occasion deviate from the booked case. Using the provided example, for instance, the surgeon may decide immediately before the procedure that they will now perform a free flap (or not perform one), or the time of day changes owing to an urgent add-on case bumping the current case. Through a quick review of the explainability, the procedure room managers can estimate how this may affect the case duration and plan case allocation and staffing accordingly for the procedures to follow. Finally, there is currently a systemic lack of trust in health care AI applications, as evidenced by several thought leaders in AI, medical ethics, and medical law [40-42]. To a significant degree, this is a result of the black box nature of most health care AI applications, seeding distrust for most health care clinicians. Providing explainability allows far greater transparency in the decision-making process and is supported by several prior studies [43-47].

Unexpected Findings

Performance metrics at each time point were ultimately similar, and many of the additional features available at later time points were dropped by the machine learning algorithm for being unimportant to model prediction. This suggests that the information provided by many of these features does not provide an overall improvement in the performance of the models and that the features with the highest importance also tend to be the ones with greater availability and at earlier time points. This provides reassurance that the model is likely to be robust within various data schemas as long as the natural language processing and feature engineering remain consistent and use electronic health record features routinely available at a majority of institutions. In addition, this can be useful for case schedulers

to fill a procedure room block efficiently, and procedure room managers can appropriately allocate resources potentially earlier.

Limitations

Despite the performance of the models, there are still a number of limitations to our approach. First, although the volume of data is high, the data as provided are relatively uniformly curated. Although this may be seen as a benefit from a data analysis perspective, it does mean that the precise data processing performed here is specific to this data structure and not necessarily to local institution data structure. The single-institution validation model aids in supporting potential generalizability, but data processing may differ by institution. Two simple solutions include using a shadow copy of local data that restructures data to the same schema or retraining of the model using local data schema. Second, the features with the highest variable importance need to be both available and reliable. Third, financial analyses related to time are beyond the scope of this study owing to multiple factors being involved, including staffing models and staffing ratios, procedure type, procedure acuity, payer status, and local policy [48,49]. Next, there may be procedures that occurred on the same patient. Ultimately, the explainability analyses suggested that patient characteristics had little contribution to make to the model performance compared with the more impactful derived features and natural language processed procedure text. In addition, the data used in this study are all from before the international COVID-19 pandemic because that is when the analysis was initially performed. The algorithm would have to be updated to include more recent postpandemic data because hospital systems are likely to have changed. Finally, all institutions in this data set are from the United States, which may limit international generalizability.

Use in Practice

We are transparent in our design and have provided the code to implement the models in [Multimedia Appendix 1](#). A code use schematic (Figure S2 in [Multimedia Appendix 1](#)) aids in understanding the relationship of processing data, updating models, and generating output. All code is available in a web-based repository [50]. First, we provide the trained models in R *H2O* format, which can be applied directly to new data (upcoming procedures) to generate predictions ([Multimedia Appendix 1](#) [refer to R Code: Making Predictions Using Created Machine Learning Models]). We provide the code needed to preprocess data, including generating derived experience features and natural language processing of procedure text ([Multimedia Appendix 1](#) [refer to R Code for Data Processing]). This preprocessing code can generate a new training set or can be applied to reformat new data for the provided models. Finally, we provide the code to generate new models or to update the existing models with more current data, including up-to-date derived experience features ([Multimedia Appendix 1](#) [refer to R Code: Training and Testing ML Models]).

For use in practice, ideally, the model will be installed and maintained locally. It can be rebuilt periodically to avoid excessive computational requirements. Time for prediction is negligible (1 s -0.3 s to +0.3 s). The model can be used as the default prediction when scheduling cases, or, if used at the time

of scheduling, it can drive alerts for procedures with scheduled times incongruent with predicted times. A recent example of a single-center prospective implementation of a similar model suggests that there is a benefit to using these methods with regard to accurate prediction of surgical times and impact on workflow [39]. However, ultimately, institutional policy will largely steer implementation; for example, many institutions do not routinely use a surgeon or scheduler estimate at procedure booking [51]. The use of this tool obviates the need for individualized input. Future studies are necessary to

prospectively validate the performance of procedure duration prediction models integrated into daily workflow for clinician and administrator use in real time.

Conclusions

We report a robust and generalizable model for the prediction of procedure duration and variability within an acceptable tolerance derived from rigorous testing of machine learning models applied to a large multicenter data set. Our findings may guide the future development of procedure room workflow implementation of procedure duration prediction models.

Acknowledgments

AB was supported by a National Institute of General Medical Sciences training grant (T32 GM008440). MM reports grants from the National Heart, Lung, and Blood Institute of the National Institutes of Health (K01-HL141701) during the conduct of the study. Support for other investigators was provided from institutional and departmental sources. Funding was provided by departmental and institutional resources at each contributing site. In addition, partial funding to support underlying electronic health record data collection into the Multicenter Perioperative Outcomes Group registry was provided by Blue Cross Blue Shield of Michigan and Blue Care Network as part of the Blue Cross Blue Shield of Michigan and Blue Care Network Value Partnerships program. Although Blue Cross Blue Shield of Michigan and Blue Care Network and the Multicenter Perioperative Outcomes Group work collaboratively, the opinions, beliefs, and viewpoints expressed by the authors do not necessarily reflect the opinions, beliefs, and viewpoints of Blue Cross Blue Shield of Michigan and Blue Care Network or its employees.

Conflicts of Interest

AB is a co-founder of Bezel Health, a company building software to measure and improve healthcare quality interventions. SS is a co-founder of Orchestra Health Inc, a digital health startup company improving care transitions. This is unrelated to the work in this study.

Multimedia Appendix 1

Supplementary tables and figures, as well as R code for data manipulation, model training, and model building. [[DOCX File, 2796 KB - ai_v2ie44909_app1.docx](#)]

References

1. Glance LG, Dutton RP, Feng C, Li Y, Lustik SJ, Dick AW. Variability in case durations for common surgical procedures. *Anesth Analg* 2018 Jun;126(6):2017-2024. [doi: [10.1213/ANE.0000000000002882](https://doi.org/10.1213/ANE.0000000000002882)] [Medline: [29517575](https://pubmed.ncbi.nlm.nih.gov/29517575/)]
2. Levine WC, Dunn PF. Optimizing operating room scheduling. *Anesthesiol Clin* 2015 Dec;33(4):697-711. [doi: [10.1016/j.anclin.2015.07.006](https://doi.org/10.1016/j.anclin.2015.07.006)] [Medline: [26610624](https://pubmed.ncbi.nlm.nih.gov/26610624/)]
3. Wu A, Huang CC, Weaver MJ, Urman RD. Use of historical surgical times to predict duration of primary total knee arthroplasty. *J Arthroplasty* 2016 Dec;31(12):2768-2772. [doi: [10.1016/j.arth.2016.05.038](https://doi.org/10.1016/j.arth.2016.05.038)] [Medline: [27396691](https://pubmed.ncbi.nlm.nih.gov/27396691/)]
4. Dexter F, Ledolter J, Tiwari V, Epstein RH. Value of a scheduled duration quantified in terms of equivalent numbers of historical cases. *Anesth Analg* 2013 Jul;117(1):205-210. [doi: [10.1213/ANE.0b013e318291d388](https://doi.org/10.1213/ANE.0b013e318291d388)] [Medline: [23733843](https://pubmed.ncbi.nlm.nih.gov/23733843/)]
5. Edelman ER, van Kuijk SM, Hamaekers AE, de Korte MJ, van Merode GG, Buhre WF. Improving the prediction of total surgical procedure time using linear regression modeling. *Front Med (Lausanne)* 2017 Jun 19;4:85 [FREE Full text] [doi: [10.3389/fmed.2017.00085](https://doi.org/10.3389/fmed.2017.00085)] [Medline: [28674693](https://pubmed.ncbi.nlm.nih.gov/28674693/)]
6. Stepaniak PS, Heij C, Mannaerts GH, de Quelerij M, de Vries G. Modeling procedure and surgical times for current procedural terminology-anesthesia-surgeon combinations and evaluation in terms of case-duration prediction and operating room efficiency: a multicenter study. *Anesth Analg* 2009 Oct;109(4):1232-1245. [doi: [10.1213/ANE.0b013e3181b5de07](https://doi.org/10.1213/ANE.0b013e3181b5de07)] [Medline: [19762753](https://pubmed.ncbi.nlm.nih.gov/19762753/)]
7. Dexter F, Macario A, O'Neill L. A strategy for deciding operating room assignments for second-shift anesthetists. *Anesth Analg* 1999 Oct;89(4):920-924. [doi: [10.1097/0000539-199910000-00019](https://doi.org/10.1097/0000539-199910000-00019)] [Medline: [10512265](https://pubmed.ncbi.nlm.nih.gov/10512265/)]
8. van Eijk RP, van Veen-Berkx E, Kazemier G, Eijkemans MJ. Effect of individual surgeons and anesthesiologists on operating room time. *Anesth Analg* 2016 Aug;123(2):445-451. [doi: [10.1213/ANE.0000000000001430](https://doi.org/10.1213/ANE.0000000000001430)] [Medline: [27308953](https://pubmed.ncbi.nlm.nih.gov/27308953/)]
9. Soh KW, Walker C, O'Sullivan M, Wallace J. Comparison of jackknife and hybrid-boost model averaging to predict surgery durations: a case study. *SN Comput Sci* 2020 Oct 01;1:316 [FREE Full text] [doi: [10.1007/s42979-020-00339-0](https://doi.org/10.1007/s42979-020-00339-0)]
10. Zhao B, Waterman RS, Urman RD, Gabriel RA. A machine learning approach to predicting case duration for robot-assisted surgery. *J Med Syst* 2019 Jan 05;43(2):32. [doi: [10.1007/s10916-018-1151-y](https://doi.org/10.1007/s10916-018-1151-y)] [Medline: [30612192](https://pubmed.ncbi.nlm.nih.gov/30612192/)]

11. Lam SS, Zaribafzadeh H, Ang BY, Webster W, Buckland D, Mantyh C, et al. Estimation of surgery durations using machine learning methods—a cross-country multi-site collaborative study. *Healthcare (Basel)* 2022 Jun 25;10(7):1191 [FREE Full text] [doi: [10.3390/healthcare10071191](https://doi.org/10.3390/healthcare10071191)] [Medline: [35885718](https://pubmed.ncbi.nlm.nih.gov/35885718/)]
12. Jiao Y, Sharma A, Ben Abdallah A, Maddox TM, Kannampallil T. Probabilistic forecasting of surgical case duration using machine learning: model development and validation. *J Am Med Inform Assoc* 2020 Dec 09;27(12):1885-1893 [FREE Full text] [doi: [10.1093/jamia/ocaa140](https://doi.org/10.1093/jamia/ocaa140)] [Medline: [33031543](https://pubmed.ncbi.nlm.nih.gov/33031543/)]
13. Tuwatananurak JP, Zadeh S, Xu X, Vacanti JA, Fulton WR, Ehrenfeld JM, et al. Machine learning can improve estimation of surgical case duration: a pilot study. *J Med Syst* 2019 Jan 17;43(3):44. [doi: [10.1007/s10916-019-1160-5](https://doi.org/10.1007/s10916-019-1160-5)] [Medline: [30656433](https://pubmed.ncbi.nlm.nih.gov/30656433/)]
14. Rozario N, Rozario D. Can machine learning optimize the efficiency of the operating room in the era of COVID-19? *Can J Surg* 2020;63(6):E527-E529 [FREE Full text] [doi: [10.1503/cjs.016520](https://doi.org/10.1503/cjs.016520)] [Medline: [33180692](https://pubmed.ncbi.nlm.nih.gov/33180692/)]
15. Bellini V, Guzzon M, Bigliardi B, Mordonini M, Filippelli S, Bignami E. Artificial intelligence: a new tool in operating room management. Role of machine learning models in operating room optimization. *J Med Syst* 2019 Dec 10;44(1):20. [doi: [10.1007/s10916-019-1512-1](https://doi.org/10.1007/s10916-019-1512-1)] [Medline: [31823034](https://pubmed.ncbi.nlm.nih.gov/31823034/)]
16. Gu on AC, Paalvast M, Meeuwse FC, Tax DM, van Dijke AP, Wauben LS, et al. Real-time estimation of surgical procedure duration. 2015 Presented at: 17th International Conference on E-health Networking, Application & Services (HealthCom); October 14-17, 2015; Boston, MA. [doi: [10.1109/HealthCom.2015.7454464](https://doi.org/10.1109/HealthCom.2015.7454464)]
17. Luo W, Phung D, Tran T, Gupta S, Rana S, Karmakar C, et al. Guidelines for developing and reporting machine learning predictive models in biomedical research: a multidisciplinary view. *J Med Internet Res* 2016 Dec 16;18(12):e323 [FREE Full text] [doi: [10.2196/jmir.5870](https://doi.org/10.2196/jmir.5870)] [Medline: [27986644](https://pubmed.ncbi.nlm.nih.gov/27986644/)]
18. Collins GS, Moons KG. Reporting of artificial intelligence prediction models. *Lancet* 2019 Apr 20;393(10181):1577-1579. [doi: [10.1016/S0140-6736\(19\)30037-6](https://doi.org/10.1016/S0140-6736(19)30037-6)] [Medline: [31007185](https://pubmed.ncbi.nlm.nih.gov/31007185/)]
19. Perioperative Clinical Research Committee review. Multicenter Perioperative Outcomes Group. URL: <https://mpog.org/general-pcrg-information/> [accessed 2021-05-21]
20. Colquhoun DA, Shanks AM, Kapeles SR, Shah N, Saager L, Vaughn MT, et al. Considerations for integration of perioperative electronic health records across institutions for research and quality improvement: the approach taken by the multicenter perioperative outcomes group. *Anesth Analg* 2020 May;130(5):1133-1146 [FREE Full text] [doi: [10.1213/ANE.0000000000004489](https://doi.org/10.1213/ANE.0000000000004489)] [Medline: [32287121](https://pubmed.ncbi.nlm.nih.gov/32287121/)]
21. Burns ML, Mathis MR, Vandervest J, Tan X, Lu B, Colquhoun DA, et al. Classification of current procedural terminology codes from electronic health record data using machine learning. *Anesthesiology* 2020 Apr;132(4):738-749 [FREE Full text] [doi: [10.1097/ALN.0000000000003150](https://doi.org/10.1097/ALN.0000000000003150)] [Medline: [32028374](https://pubmed.ncbi.nlm.nih.gov/32028374/)]
22. Sun E, Mello MM, Rishel CA, Vaughn MT, Khetarpal S, Saager L, Multicenter Perioperative Outcomes Group (MPOG). Association of overlapping surgery with perioperative outcomes. *JAMA* 2019 Feb 26;321(8):762-772 [FREE Full text] [doi: [10.1001/jama.2019.0711](https://doi.org/10.1001/jama.2019.0711)] [Medline: [30806696](https://pubmed.ncbi.nlm.nih.gov/30806696/)]
23. Lee LO, Bateman BT, Khetarpal S, Klumpner TT, Housey M, Aziz MF, Multicenter Perioperative Outcomes Group Investigators. Risk of epidural hematoma after neuraxial techniques in thrombocytopenic parturients: a report from the multicenter perioperative outcomes group. *Anesthesiology* 2017 Jun;126(6):1053-1063 [FREE Full text] [doi: [10.1097/ALN.0000000000001630](https://doi.org/10.1097/ALN.0000000000001630)] [Medline: [28383323](https://pubmed.ncbi.nlm.nih.gov/28383323/)]
24. Aziz MF, Brambrink AM, Healy DW, Willett AW, Shanks A, Tremper T, et al. Success of intubation rescue techniques after failed direct laryngoscopy in adults: a retrospective comparative analysis from the multicenter perioperative outcomes group. *Anesthesiology* 2016 Oct;125(4):656-666 [FREE Full text] [doi: [10.1097/ALN.0000000000001267](https://doi.org/10.1097/ALN.0000000000001267)] [Medline: [27483124](https://pubmed.ncbi.nlm.nih.gov/27483124/)]
25. Strum DP, Sampson AR, May JH, Vargas LG. Surgeon and type of anesthesia predict variability in surgical procedure times. *Anesthesiology* 2000 May;92(5):1454-1466 [FREE Full text] [doi: [10.1097/0000542-200005000-00036](https://doi.org/10.1097/0000542-200005000-00036)] [Medline: [10781292](https://pubmed.ncbi.nlm.nih.gov/10781292/)]
26. Bartek MA, Saxena RC, Solomon S, Fong CT, Behara LD, Venigandla R, et al. Improving operating room efficiency: machine learning approach to predict case-time duration. *J Am Coll Surg* 2019 Oct;229(4):346-54.e3 [FREE Full text] [doi: [10.1016/j.jamcollsurg.2019.05.029](https://doi.org/10.1016/j.jamcollsurg.2019.05.029)] [Medline: [31310851](https://pubmed.ncbi.nlm.nih.gov/31310851/)]
27. Althnian A, AlSaeed D, Al-Baity H, Samha A, Dris AB, Alzakari N, et al. Impact of dataset size on classification performance: an empirical evaluation in the medical domain. *Appl Sci* 2021 Jan 15;11(2):796. [doi: [10.3390/app11020796](https://doi.org/10.3390/app11020796)]
28. van der Ploeg T, Austin PC, Steyerberg EW. Modern modelling techniques are data hungry: a simulation study for predicting dichotomous endpoints. *BMC Med Res Methodol* 2014 Dec 22;14:137 [FREE Full text] [doi: [10.1186/1471-2288-14-137](https://doi.org/10.1186/1471-2288-14-137)] [Medline: [25532820](https://pubmed.ncbi.nlm.nih.gov/25532820/)]
29. Aiello S, Eckstrand E, Fu A, Landry M, Aboyou P. Machine learning with R and H2O. H2O. 2015 Dec. URL: http://h2o-release.s3.amazonaws.com/h2o/master/3283/docs-website/h2o-docs/booklets/R_Vignette.pdf [accessed 2023-08-16]
30. Goodfellow I, Bengio Y, Courville A. Deep Learning. Cambridge, MA: MIT Press; Nov 18, 2016.
31. Click C, Malohlava M, Roark H, Parmar V, Lanford J. Gradient boosting machine with H2O. H2O. 2015 Aug. URL: https://h2o-release.s3.amazonaws.com/h2o/master/3147/docs-website/h2o-docs/booklets/GBM_Vignette.pdf [accessed 2023-08-16]

32. LeDell E, Poirier S. H2O AutoML: scalable automatic machine learning. 2020 Presented at: 7th ICML Workshop on Automated Machine Learning; July 17-18, 2020; Vienna, Austria. [doi: [10.1007/s10994-022-06262-0](https://doi.org/10.1007/s10994-022-06262-0)]
33. Candel A, Parmar V, LeDell E, Arora A. Deep learning with H2O. H2O. 2016 Oct. URL: <https://www.webpages.uidaho.edu/~stevel/504/Deep%20Learning%20with%20H2O.pdf> [accessed 2023-08-16]
34. Pandit JJ, Tavare A. Using mean duration and variation of procedure times to plan a list of surgical operations to fit into the scheduled list time. *Eur J Anaesthesiol* 2011 Jul;28(7):493-501. [doi: [10.1097/EJA.0b013e3283446b9c](https://doi.org/10.1097/EJA.0b013e3283446b9c)] [Medline: [21623186](https://pubmed.ncbi.nlm.nih.gov/21623186/)]
35. Lundberg SM, Erion G, Chen H, DeGrave A, Prutkin JM, Nair B, et al. From local explanations to global understanding with explainable AI for trees. *Nat Mach Intell* 2020 Jan;2(1):56-67 [FREE Full text] [doi: [10.1038/s42256-019-0138-9](https://doi.org/10.1038/s42256-019-0138-9)] [Medline: [32607472](https://pubmed.ncbi.nlm.nih.gov/32607472/)]
36. Altmann A, Tološi L, Sander O, Lengauer T. Permutation importance: a corrected feature importance measure. *Bioinformatics* 2010 May 15;26(10):1340-1347. [doi: [10.1093/bioinformatics/btq134](https://doi.org/10.1093/bioinformatics/btq134)] [Medline: [20385727](https://pubmed.ncbi.nlm.nih.gov/20385727/)]
37. Knapič S, Malhi A, Saluja R, Främling K. Explainable artificial intelligence for human decision-support system in medical domain. arXiv Preprint posted online May 5, 2021. [FREE Full text] [doi: [10.3390/make3030037](https://doi.org/10.3390/make3030037)]
38. Lundberg SM, Nair B, Vavilala MS, Horibe M, Eisses MJ, Adams T, et al. Explainable machine-learning predictions for the prevention of hypoxaemia during surgery. *Nat Biomed Eng* 2018 Oct;2(10):749-760. [doi: [10.1038/s41551-018-0304-0](https://doi.org/10.1038/s41551-018-0304-0)] [Medline: [31001455](https://pubmed.ncbi.nlm.nih.gov/31001455/)]
39. Strömblad CT, Baxter-King RG, Meisami A, Yee SJ, Levine MR, Ostrovsky A, et al. Effect of a predictive model on planned surgical duration accuracy, patient wait time, and use of presurgical resources: a randomized clinical trial. *JAMA Surg* 2021 Apr 01;156(4):315-321 [FREE Full text] [doi: [10.1001/jamasurg.2020.6361](https://doi.org/10.1001/jamasurg.2020.6361)] [Medline: [33502448](https://pubmed.ncbi.nlm.nih.gov/33502448/)]
40. Gille F, Jobin A, Ienca M. What we talk about when we talk about trust: theory of trust for AI in healthcare. *Intell Based Med* 2020 Nov;1-2:100001. [doi: [10.1016/j.ibmed.2020.100001](https://doi.org/10.1016/j.ibmed.2020.100001)]
41. Hatherley JJ. Limits of trust in medical AI. *J Med Ethics* 2020 Jul;46(7):478-481. [doi: [10.1136/medethics-2019-105935](https://doi.org/10.1136/medethics-2019-105935)] [Medline: [32220870](https://pubmed.ncbi.nlm.nih.gov/32220870/)]
42. Durán JM, Jongsma KR. Who is afraid of black box algorithms? On the epistemological and ethical basis of trust in medical AI. *J Med Ethics* 2021 Mar 18:medethics-2020-106820. [doi: [10.1136/medethics-2020-106820](https://doi.org/10.1136/medethics-2020-106820)] [Medline: [33737318](https://pubmed.ncbi.nlm.nih.gov/33737318/)]
43. Kundu S. AI in medicine must be explainable. *Nat Med* 2021 Aug 29;27(8):1328. [doi: [10.1038/s41591-021-01461-z](https://doi.org/10.1038/s41591-021-01461-z)] [Medline: [34326551](https://pubmed.ncbi.nlm.nih.gov/34326551/)]
44. Liu CF, Chen ZC, Kuo SC, Lin TC. Does AI explainability affect physicians' intention to use AI? *Int J Med Inform* 2022 Dec;168:104884. [doi: [10.1016/j.ijmedinf.2022.104884](https://doi.org/10.1016/j.ijmedinf.2022.104884)] [Medline: [36228415](https://pubmed.ncbi.nlm.nih.gov/36228415/)]
45. Diprose WK, Buist N, Hua N, Thurier Q, Shand G, Robinson R. Physician understanding, explainability, and trust in a hypothetical machine learning risk calculator. *J Am Med Inform Assoc* 2020 Apr 01;27(4):592-600 [FREE Full text] [doi: [10.1093/jamia/ocz229](https://doi.org/10.1093/jamia/ocz229)] [Medline: [32106285](https://pubmed.ncbi.nlm.nih.gov/32106285/)]
46. Lötsch J, Kringel D, Ultsch A. Explainable Artificial Intelligence (XAI) in biomedicine: making AI decisions trustworthy for physicians and patients. *BioMedInformatics* 2021 Dec 22;2(1):1-17. [doi: [10.3390/biomedinformatics2010001](https://doi.org/10.3390/biomedinformatics2010001)]
47. Amann J, Blasimme A, Vayena E, Frey D, Madai VI, Precise4Q consortium. Explainability for artificial intelligence in healthcare: a multidisciplinary perspective. *BMC Med Inform Decis Mak* 2020 Nov 30;20(1):310 [FREE Full text] [doi: [10.1186/s12911-020-01332-6](https://doi.org/10.1186/s12911-020-01332-6)] [Medline: [33256715](https://pubmed.ncbi.nlm.nih.gov/33256715/)]
48. Xiang W, Li C. Surgery scheduling optimization considering real life constraints and comprehensive operation cost of operating room. *Technol Health Care* 2015;23(5):605-617. [doi: [10.3233/THC-151017](https://doi.org/10.3233/THC-151017)] [Medline: [26410121](https://pubmed.ncbi.nlm.nih.gov/26410121/)]
49. Chapman WC, Luo X, Doyle M, Khan A, Chapman WC, Kangrga I, et al. Time is money: can punctuality decrease operating room cost? *J Am Coll Surg* 2020 Feb;230(2):182-9.e4. [doi: [10.1016/j.jamcollsurg.2019.10.017](https://doi.org/10.1016/j.jamcollsurg.2019.10.017)] [Medline: [31843690](https://pubmed.ncbi.nlm.nih.gov/31843690/)]
50. SuperProcLengthEngine. GitHub. URL: <http://github.com/skendalem/SuperProcLengthEngine> [accessed 2022-10-10]
51. Wright IH, Kooperberg C, Bonar BA, Bashein G. Statistical modeling to predict elective surgery time. Comparison with a computer scheduling system and surgeon-provided estimates. *Anesthesiology* 1996 Dec;85(6):1235-1245 [FREE Full text] [doi: [10.1097/0000542-199612000-00003](https://doi.org/10.1097/0000542-199612000-00003)] [Medline: [8968169](https://pubmed.ncbi.nlm.nih.gov/8968169/)]

Abbreviations

- AI:** artificial intelligence
- CPT:** current procedural terminology
- MAE:** mean absolute error
- MAPE:** mean absolute percentage error
- MPOG:** Multicenter Perioperative Outcomes Group
- NYU:** New York University
- RMSE:** root mean square error
- SHAP:** Shapley additive explanations

Edited by K El Emam; submitted 09.12.22; peer-reviewed by C Valderrama Cuadros, JA Benítez-Andrades, E Bignami, PF Chen, L Nakayama; comments to author 06.02.23; revised version received 14.06.23; accepted 02.07.23; published 08.09.23.

Please cite as:

Kendale S, Bishara A, Burns M, Solomon S, Corriere M, Mathis M

Machine Learning for the Prediction of Procedural Case Durations Developed Using a Large Multicenter Database: Algorithm Development and Validation Study

JMIR AI 2023;2:e44909

URL: <https://ai.jmir.org/2023/1/e44909>

doi: [10.2196/44909](https://doi.org/10.2196/44909)

PMID:

©Samir Kendale, Andrew Bishara, Michael Burns, Stuart Solomon, Matthew Corriere, Michael Mathis. Originally published in JMIR AI (<https://ai.jmir.org>), 08.09.2023. This is an open-access article distributed under the terms of the Creative Commons Attribution License (<https://creativecommons.org/licenses/by/4.0/>), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR AI, is properly cited. The complete bibliographic information, a link to the original publication on <https://www.ai.jmir.org/>, as well as this copyright and license information must be included.

Original Paper

Association of Health Care Work With Anxiety and Depression During the COVID-19 Pandemic: Structural Topic Modeling Study

Matteo Malgaroli¹, PhD; Emily Tseng², MSc; Thomas D Hull³, PhD; Emma Jennings¹, MSc; Tanzeem K Choudhury², PhD; Naomi M Simon¹, MSc, MD

¹Department of Psychiatry, Grossman School of Medicine, New York University, New York, NY, United States

²Ann S Bowers College of Computing and Information Science, Cornell University, Ithaca, NY, United States

³Research and Development, Talkspace, New York, NY, United States

Corresponding Author:

Matteo Malgaroli, PhD

Department of Psychiatry

Grossman School of Medicine

New York University

1 Park Avenue

8th Floor

New York, NY, 10016

United States

Phone: 1 6467544030

Email: matteo.malgaroli@nyulangone.org

Abstract

Background: Stressors for health care workers (HCWs) during the COVID-19 pandemic have been manifold, with high levels of depression and anxiety alongside gaps in care. Identifying the factors most tied to HCWs' psychological challenges is crucial to addressing HCWs' mental health needs effectively, now and for future large-scale events.

Objective: In this study, we used natural language processing methods to examine deidentified psychotherapy transcripts from telemedicine treatment during the initial wave of COVID-19 in the United States. Psychotherapy was delivered by licensed therapists while HCWs were managing increased clinical demands and elevated hospitalization rates, in addition to population-level social distancing measures and infection risks. Our goal was to identify specific concerns emerging in treatment for HCWs and to compare differences with matched non-HCW patients from the general population.

Methods: We conducted a case-control study with a sample of 820 HCWs and 820 non-HCW matched controls who received digitally delivered psychotherapy in 49 US states in the spring of 2020 during the first US wave of the COVID-19 pandemic. Depression was measured during the initial assessment using the Patient Health Questionnaire-9, and anxiety was measured using the General Anxiety Disorder-7 questionnaire. Structural topic models (STMs) were used to determine treatment topics from deidentified transcripts from the first 3 weeks of treatment. STM effect estimators were also used to examine topic prevalence in patients with moderate to severe anxiety and depression.

Results: The median treatment enrollment date was April 15, 2020 (IQR March 31 to April 27, 2020) for HCWs and April 19, 2020 (IQR April 5 to April 27, 2020) for matched controls. STM analysis of deidentified transcripts identified 4 treatment topics centered on health care and 5 on mental health for HCWs. For controls, 3 STM topics on pandemic-related disruptions and 5 on mental health were identified. Several STM treatment topics were significantly associated with moderate to severe anxiety and depression, including working on the hospital unit (topic prevalence 0.035, 95% CI 0.022-0.048; $P < .001$), mood disturbances (prevalence 0.014, 95% CI 0.002-0.026; $P = .03$), and sleep disturbances (prevalence 0.016, 95% CI 0.002-0.030; $P = .02$). No significant associations emerged between pandemic-related topics and moderate to severe anxiety and depression for non-HCW controls.

Conclusions: The study provides large-scale quantitative evidence that during the initial wave of the COVID-19 pandemic, HCWs faced unique work-related challenges and stressors associated with anxiety and depression, which required dedicated treatment efforts. The study further demonstrates how natural language processing methods have the potential to surface clinically relevant markers of distress while preserving patient privacy.

KEYWORDS

depression; anxiety; health care workers; COVID-19; natural language processing; topic modeling; stressor; mental health; treatment; psychotherapy; digital health

Introduction

During the COVID-19 pandemic, health care workers (HCWs) faced mounting stress as they cared for patients experiencing a disease that, to date, has infected 538 million globally and 85 million in the United States alone [1]. Surges in US infection rates forced hospitals to operate at greater than 100% capacity [2], with COVID-19 hospitalizations in 18 states exceeding 10% of all available beds and 7 states operating at more than 15% overcapacity [3]. As a result, HCWs faced overwhelming workloads, longer hours, increased personal infection risk, equipment shortages, sleep disruption, and at times the need to make ethically challenging decisions, such as rationing care for patients [4-8]. This increased burden on HCWs was further aggravated by the loss of social support due to quarantine policies and the fear of infecting family and friends [6,7,9].

The well-being of HCWs is the foundation of a well-functioning health care system [10-12]. Prior to the pandemic, HCWs already faced higher rates of anxiety, depression [13], and suicidal ideation [14] compared to the general population [13,15]. The sudden increase in professional and personal stress due to COVID-19 put HCWs, an already vulnerable population with barriers to treatment access [16], at further risk for developing symptoms of anxiety and depression [5,6,9]. Prior studies have linked depression and anxiety in HCWs to decreased patient safety and increased medical errors [17-20]. Given the adverse consequences of psychological stress for HCWs and their patients, it is crucial to better understand the core concerns associated with mental health symptoms such as anxiety and depression in HCWs, especially during periods of acute stress like COVID-19 surges. It is especially crucial to study these concerns in ways that preserve the privacy and anonymity of HCWs, given the professional stigma reported by some health care providers who seek mental health treatment [21-23].

Hastened by the pandemic, recent advances have been made in developing and disseminating digital mental health interventions to address acute and long-term treatment barriers, including mobile apps and telehealth platforms connecting patients to mental health providers [24]. Such interventions offer a unique opportunity for understanding and addressing the mental health concerns of vulnerable populations like HCWs. Despite their potential, little research has examined the adoption of digital health treatment by HCWs during COVID-19.

In addition to providing flexible options for clinical engagement, digital treatment delivery enables the automatic collection of large amounts of treatment data, which in turn can be analyzed in an aggregated and deidentified fashion using machine learning (ML) methods. Researchers in digital psychiatry and ubiquitous computing have used ML to develop passive measures for mental health concerns, which can be refined into clinically

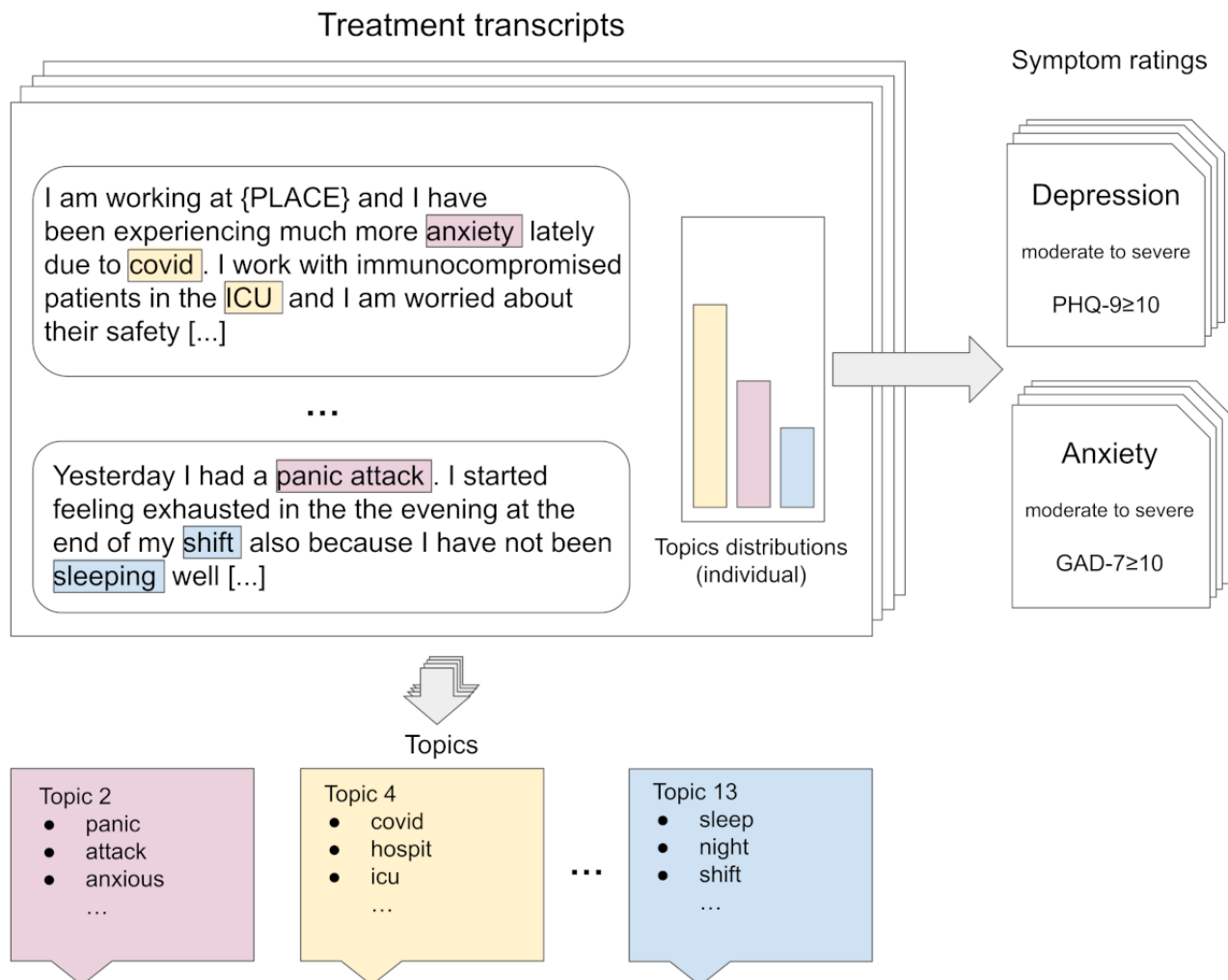
relevant markers for symptom severity and embedded into treatment pathways [25-27]. ML-based natural language processing (NLP) holds particular promise for the study of mental health concerns, as it allows the study of verbal expressions of distress at scale, capturing clinically relevant linguistic features from unstructured text as the patient-therapist interaction unfolds [28]. Of particular interest in the study of psychotherapy transcripts is topic modeling, an unsupervised NLP method to parse semantic structures (or topics) from large corpora of text without the need for line-by-line annotation [29]. Topic modeling has been used to generate knowledge in multiple areas of science [30], and previous uses of topic modeling in mental health include the detection of depression [31] and anxiety [32], also in the context of the COVID-19 pandemic [33]. In brief, topic modeling imagines that every document within a corpus contains a mixture of corpus-wide distributions of words within a fixed vocabulary. Topic modeling algorithms seek to find the topics that best characterize a given corpus across documents, as a means to understand the core content of potentially difficult texts (such as therapy transcripts) at scale [28]. Structural topic models (STMs) also enable the study of covariates in their influence on topic prevalence, or the proportion of a document associated with a topic, and topical content, or the distribution of words used within a topic. Compared to lexicon-based methods, topic modeling allows assessing context-specific language (such as medical terminology) within the corpus of transcripts. Compared to embeddings, which capture semantic similarity at the word or sentence level, topic modeling can also uncover broader thematic associations across transcripts, to then group individuals based on topical themes emerging from the transcripts. For collections of texts like transcripts of psychotherapy sessions, topic modeling also offers the potential to be more privacy preserving: topic models process text into distributions of keywords and enable researchers to study the semantic content of sensitive therapy transcripts while preserving treatment-seekers' privacy by minimizing exposure of personally identifiable or sensitive data. Topic modeling can provide empirical insights into the stressors experienced by medical providers during this highly stressful period of the pandemic. Moreover, by linking specific concerns identified via topic modeling with depression and anxiety symptoms measured with validated scales, linguistic features can be developed to serve as passive computational markers [34] of distress, with the potential to highlight areas of risk or need for clinical attention. Identifying the most disruptive risk factors for HCWs would support improvements in treatment planning and inform the selection of mental health resources for HCWs now, during future COVID-19 waves, or other widespread epidemics.

In this study, we examined deidentified treatment transcripts from 820 HCWs and 820 matched controls who received digitally delivered psychotherapy from licensed therapists during the first US surge of the COVID-19 pandemic, between March

and July 2020. Our aim was to identify the unique treatment needs of HCWs characterized as treatment topics compared to non-HCW matched controls. We analyzed transcripts using STM to assess the topic in the deidentified transcripts and their associations with symptom levels. Specifically, we used topic

modeling to analyze therapeutic conversations during the first 3 weeks of treatment in HCW and controls and identified emerging topics in a privacy-preserving fashion. We also assessed the association of these topics with moderate to severe levels of anxiety and depression (Figure 1).

Figure 1. Schematic overview of topic modeling with fictitious example of a transcript. Topics are generated across the full transcript corpora. Individual topic distributions are then associated with their respective symptom ratings. GAD-7: General Anxiety Disorder Scale-7; ICU: intensive care unit; PHQ-9: patient health questionnaire-9.



Methods

Participants and Setting

Our sample consisted of self-referred HCWs from the United States seeking digitally delivered psychotherapy in spring 2020, amidst the first US surge of COVID-19 hospitalizations. HCWs were defined as health care and medical providers (eg, physicians, nurses, residents, emergency medical service providers, and social workers) with an active National Provider Identifier (NPI) profile at the time of treatment. Services were donated by a telehealth platform [35] as part of an initiative to provide 1 month of free treatment to essential HCWs. Eligibility was verified by the platform through employment and NPI verification. In order to distinguish health care-specific and general population stressors related to COVID-19, we included a matched control sample of non-HCWs from the general population seeking the same treatment service as the HCW

sample in spring 2020. Non-HCW patients accessed the platform through employee assistance programs, self-referral, and as benefits through individual insurance. From this outpatient pool, a control sample was matched to HCWs based on demographics, symptom scores, US state of residency, and treatment start date. Control matching was performed algorithmically, and matching procedures are described in [Multimedia Appendix 1](#) [5,29,36-49].

Before starting treatment, HCWs and controls received a primary ICD-10 diagnosis based on a standardized intake evaluation by a licensed clinician to identify presenting complaints and treatment history. Following the intake, HCWs and controls were matched to a licensed therapist and received psychotherapy through messages exchanged using a HIPAA (Health Insurance Portability and Accountability Act)-compliant interface for smartphones and computers. The inclusion criteria were (1) living in the United States, (2) being an English speaker, and

(3) having regular internet or cellphone access (to access the digitally delivered treatment). Exclusion criteria for both samples were (1) any condition deemed by the intake clinician to require hospitalization; (2) suicidal thoughts or behavior sufficient to be marked a yes on any of questions 3 through 6 (at least thoughts about a potential suicide method) on the Columbia Suicide Severity Rating Scale Lifetime-Recent [36]; (3) current or past diagnoses of bipolar disorder, substance use disorders, schizophrenia spectrum disorders, or psychotic disorders; (4) patients who did not have complete baseline symptom measures; and (5) patients who did not have treatment transcripts available. Last, as exclusion criterion 6, during matching procedures, we excluded from the control group any health care professional. An overview and schematic of the sampling procedure in this study are reported in [Multimedia Appendix 1](#) [5,29,36-49].

The final sample consisted of 820 HCWs and 820 matched controls. The median treatment start date for HCWs was April 15, 2020 (IQR March 31 to April 27, 2020). For the matched control group, the median treatment start date was April 19, 2020 (IQR April 5 to April 27, 2020).

Data Sources and Measures

Transcripts

Psychotherapy treatment transcripts consisted of deidentified messages between patients and their therapists with their corresponding timestamp (ie, date and time of delivery) in masked form for the author role of the text. All transcript data were deidentified using an algorithm to scrub out any personal identifiers, proper nouns, locations, dates, and other potential identifiers. Transcripts were truncated to include only messages sent by patients from the initial intake to their first outcome survey, typically 3 weeks after treatment initiation. HCWs and control transcripts were both preprocessed for analysis: numbers, punctuation, stopwords, and anonymization terms (eg, "{NAME}") were removed; the remaining words were stemmed and converted to their root form (eg, *computing* was changed to *comput*). The "vocabulary" of unique words across the preprocessed transcripts was then made more tractable by removing words that occurred in less than 50 documents and then removing documents that contained no words. The final HCW corpus contained 820 therapy transcripts and a vocabulary of 1208 unique terms across 225,219 tokens. The final control corpus contained 820 transcripts and a vocabulary of 1259 unique terms across 217,321 tokens.

HCW Occupations

NPI information for HCWs in our study was not available as data due to privacy reasons. To assess the distribution of specific health care professions in the HCW sample anonymously, we developed a heuristic classification algorithm. The algorithm detected instances in the transcripts where patients self-identified as HCWs or spoke about their professional roles. Code, heuristics, and accuracy metrics of the heuristic classification algorithm are further reported in [Multimedia Appendices 1](#) [5,29,36-49] and 3.

Psychiatric Symptom Measures

Depression symptoms were measured at the beginning of treatment using the Patient Health Questionnaire-9 (PHQ-9) [50], and anxiety symptoms were measured using the General Anxiety Disorder Scale-7 (GAD-7) [51]. The PHQ-9 assesses for depressive symptoms over the past 2 weeks on a 4-point Likert scale (0="not at all" to 3="nearly every day"), with a total maximum score of 27. The GAD-7 examines symptoms of anxiety over the past 2 weeks on a 4-point Likert scale (0="not at all" to 3="nearly every day"), with a total maximum score of 21. A score of 10 or more on the PHQ-9 identifies the presence of clinically significant moderate-to-severe depression [50]; a score of 10 or more on the GAD-7 identifies the presence of clinically significant moderate-to-severe anxiety [51].

Data Analysis

Treatment Topic Identification

All analyses were conducted in Python (version 3.9.9) and in R (version 4.1.2) [37], using the package *stm* [38] for topic modeling. Additional model specifications, diagnostic analyses, model selection procedures, and code for all analyses are reported in [Multimedia Appendices 1](#) [5,29,36-49] and 2.

STMs were used to identify topics in the HCW and matched control corpora. We used a mixed statistical and human validation process to select topics ($K=30$) for analysis in both HCW and control data sets (see [Multimedia Appendix 1](#) [5,29,36-49] for full details). After identifying these topics, results from the STM were manually coded to characterize their relevance to one of three areas of interest: (1) mental health, (2) COVID-19 pandemic-related disruptions, and (3) health care. Classification of relevant topics was determined through the consensus of a panel of experts consisting of 2 doctoral-level clinical psychologists (MM and TDH), 1 psychiatrist (NMS), and 1 NLP researcher (ET). Topics were examined to understand their content based on their most characteristic words, determined using the harmonic mean of word frequency and exclusivity across topics [39].

Topics and Clinical Levels of Depression and Anxiety

We used STM effect estimators to study the association between topics discussed by each patient with moderate to severe depression or anxiety. Specifically, we ran logistic-normal generalized linear models examining the association between the prevalence of relevant topics in a patient's transcripts and their binarized psychopathology score, with GAD-7 or PHQ-9 symptom scores ≥ 10 classed as moderate-to-severe anxiety or depression ([Figure 1](#)). The study combined PHQ-9 ≥ 10 or GAD-7 ≥ 10 cutoffs to account for the high prevalence of anxiety and depression comorbidity [52]. To estimate the parameters of the generalized linear models, we used a global approximation to the average covariance matrix governing the variational posterior (vs a per-document approximation that was less computationally tractable). Topic prevalence for a particular topic is contrasted for 2 groups within a categorical covariate (none-to-mild vs moderate-to-severe symptoms). For each data set, all topics ($K=30$) were modeled and reported in [Multimedia Appendix 1](#) [5,29,36-49]; here, we report only those topics manually characterized as relevant.

Ethics Approval

All patients and clinicians gave informed consent to use their data in a deidentified, aggregated format for research purposes as part of the user agreement before they began using the platform. Study procedures were approved by the Cornell University Institutional Review Board (2004009578).

Results

Sample Characteristics

HCW patients (n=820, [Table 1](#)) modally identified as female (746/820, 91%). The mean age in the sample was 31.3 (SD 5.7) years. They were distributed across the United States, with the largest concentrations in New York State (114/820, 13.1%) and California (107/820, 13%). [Figure 2](#) reports the distribution of professions in the HCWs identified from the transcripts, with nurses (414/820, 50.5%) and physicians (148/820, 18.1%) being the most frequent health care occupations. For 289 HCWs

(35.2%), this was reportedly the first psychotherapy treatment experience. Primary diagnoses given by intake clinicians for HCWs included anxiety disorders (463/820, 56.5%), of which 100 were generalized anxiety disorders (12.2%). Trauma- and stressor-related disorders (275/820, 35.5%) were next most common, with adjustment disorders as the modal diagnosis (219/820, 26.7%) in this category. Finally, depressive disorders (67/820, 8.2%) were least common and included 45 diagnosed with major depressive disorder (5.5%). Based on PHQ-9 and GAD-7 cutoffs, the prevalence of moderate to severe depression in the HCW sample at the beginning of treatment was 43.9% (n=360) and moderate to severe anxiety was 68.5% (n=562). A total of 601 (73.3%) HCWs had either moderate to severe anxiety or depression at baseline. In the matched control sample, 560 (68.3%) had moderate to severe anxiety and 408 (49.8%) had moderate to severe depression, with 601 (73.3%) having either moderate to severe anxiety or depression at baseline. Characteristics for the sample of matched controls (n=820) are reported in [Table 1](#).

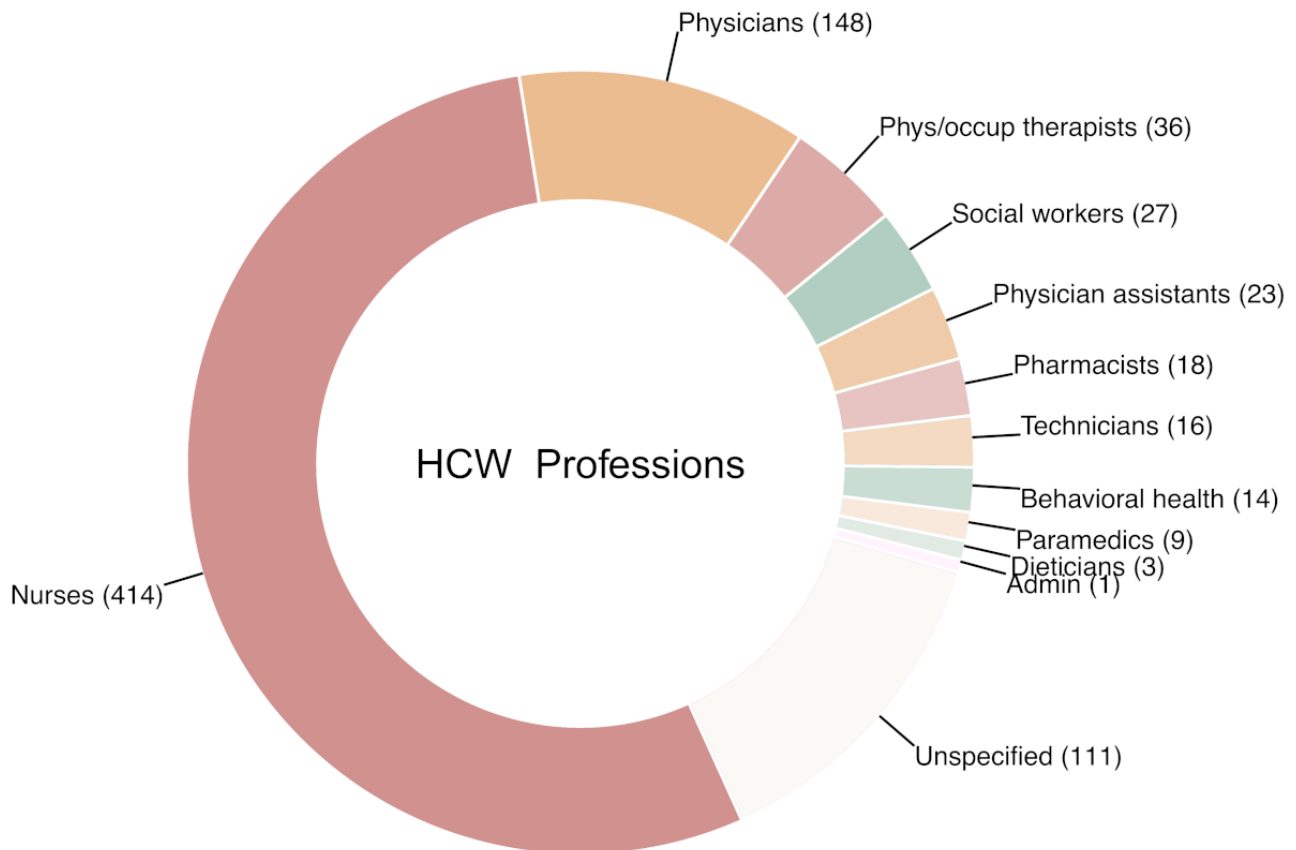
Table 1. Demographic and clinical characteristics of health care worker sample (n=820) and matched control sample (n=820).

Variable	Health care workers	Matched controls
Age (years), mean (SD)	31.3 (5.7)	32 (6.6)
Diagnosis, n (%)		
Anxiety disorders	463 (56.5)	429 (52.3)
Trauma- and stressor-related disorders	275 (35.5)	137 (16.7)
Depressive disorders	67 (8.2)	225 (27.4)
Other disorders	15 (1.8)	29 (3.5)
Gender, n (%)		
Female	746 (91)	682 (83.2)
Male	69 (8.4)	125 (15.2)
Other	5 (0.6)	13 (1.6)
State, n (%)		
California	114 (13.1)	132 (16.1)
New York	107 (13)	108 (13.2)
Florida	55 (6.7)	56 (6.8)
Texas	48 (5.9)	38 (4.6)
Illinois	45 (5.5)	33 (4)
Massachusetts	41 (5)	34 (4.2)
Pennsylvania	35 (4.3)	34 (4.2)
North Carolina	30 (3.7)	28 (3.4)
New Jersey	30 (3.7)	30 (3.7)
Washington	28 (3.4)	27 (3.9)
Other US states	287 (35)	300 (36.6)
Anxiety		
GAD-7 ^a score, mean (SD)	12.4 (4.9)	12.3 (5.1)
Moderate to severe (GAD-7 ≥ 10), n (%)	562 (68.5)	560 (68.3)
Depression		
PHQ-9 ^b score, mean (SD)	9.4 (5.7)	10 (5.8)
Moderate to severe (PHQ-9 ≥ 10), n (%)	360 (43.9)	408 (49.8)
Treatment		
First experience (Yes), n (%)	289 (35.2)	272 (33.2)
Start date (month/day/year), median (IQR)	04/15/2020 (03/31/2020-04/27/2020)	04/19/2020 (4/5/2020-4/27/2020)

^aGAD-7: General Anxiety Disorder Scale-7.

^bPHQ-9: Patient Health Questionnaire-9.

Figure 2. Algorithmically identified distribution of medical professions in the HCW sample (n=820). HCW: health care worker.



Treatment Topics

STM of psychotherapy transcripts identified 30 conversational themes for HCWs and 30 topics for non-HCW controls. Inspection of the topics showed a cluster of themes relevant to mental health and a cluster relevant to health care and the pandemic (Table 2). All topics emerging from the transcripts are reported in Figure 3.

HCWs discussed 4 topics related to practicing medicine. Examination of the most frequent words exclusive to HCWs indicated treatment topics focused on (1) virus-related fears (topic H3: *covid, worker, healthcar*), (2) working on the hospital floor and intensive care units (H4: *unit, hospit, icu*), (3) patients and masks (H16: *patient, mask, test*), and (4) health care roles including resident and attending (H29: *resid, remain, attend*). In contrast, therapy transcripts from controls contained only 1

topic about the COVID-19 pandemic (C25: *pandem, concern, anxiety*) and 1 occupational-related topic (C27: *team, manag, boss*).

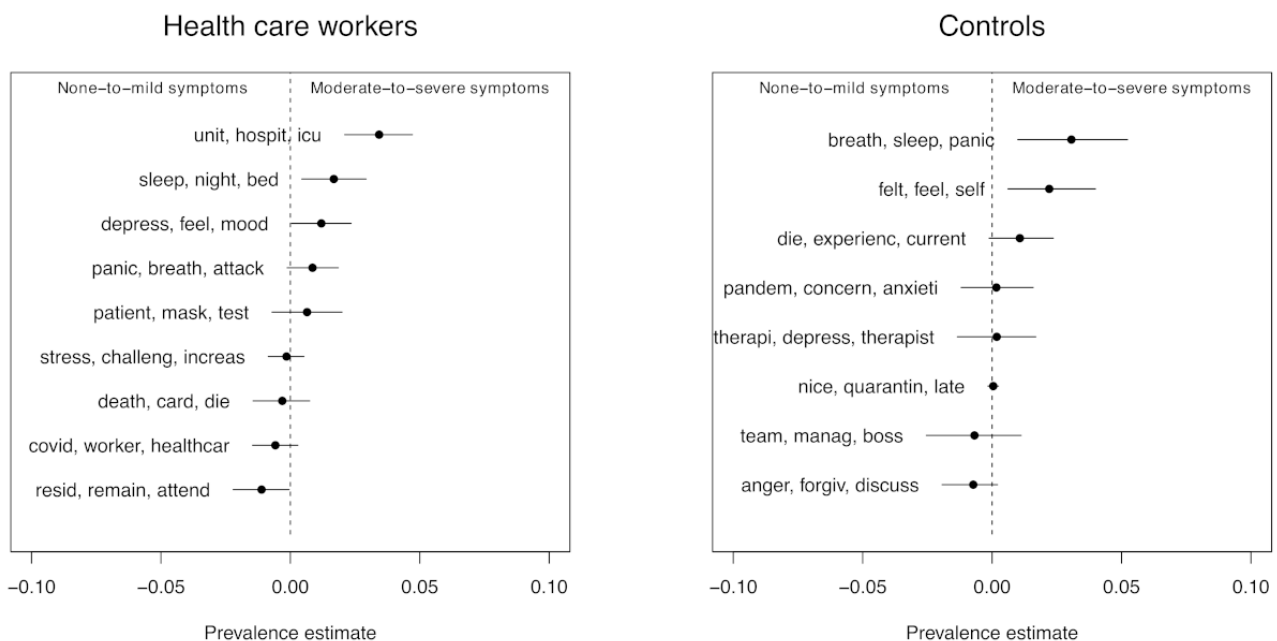
HCWs and controls each discussed 5 topics with their therapist related to their mental health, endorsing panic attacks (HCW H2: *panic, breath, attack*; control C21: *breath, sleep, panic*), affective disturbances (HCW H15: *depress, feel, mood*; control C16: *felt, feel, self*), and grief (HCW H30: *death, card, die*; control C19: *die, experienc, current*). HCWs also endorsed sleep disturbances (H13: *sleep, night, bed*), and stress (H21: *stress, challeng, increase*). Among health care and mental health topics, HCWs most frequently discussed sleep disturbances (H13: *sleep, night, bed*) and the hospital floor (H4: *unit, hospit, icu*). Multimedia Appendix 1 [5,29,36-49] reports the proportions of all topics in the HCWs and control transcripts.

Table 2. Psychotherapy topics referencing mental health, health care, and COVID-19 in health care workers and matched controls.

Sample, category, and topic	Top 10 terms (frequency and exclusivity) ^a
Health care workers (n=820)	
Health care	
H3	covid, worker, healthcar, hospit, patient, physician, current, week, promot, doctor
H4	unit, hospit, icu, nurs, virus, news, sick, covid, safe, fear
H16	patient, mask, test, shift, unit, wear, staff, icu, ppe, coronavirus
H29	resid, remain, attend, program, becom, answer, clinic, mayb, mean, studi
Mental health	
H2	panic, breath, attack, symptom, anxious, anxieti, exercis, chest, tool, calm
H13	sleep, night, bed, shift, asleep, wake, usual, fall, morn, relax
H15	depress, feel, mood, anyth, suicid, quarantin, sad, episod, sometim, hard
H21	stress, challeng, increas, relief, level, team, stressor, overal, focus, line
H30	death, card, die, grief, code, credit, pass, deal, charg, enter
Matched controls (n=820)	
Pandemic disruptions	
C25	pandem, concern, anxieti, situat, cope, corona, group, relat, social, extrem
C11	nice, quarantin, late, gym, enjoy, crazi, glad, weather, excit, heavi
C27	team, manag, boss, project, task, routin, offic, work, cowork, hour
Mental health	
C21	breath, sleep, panic, sick, attack, night, anxious, anxieti, worri, calm
C16	felt, feel, self, negat, anxious, thought, sad, bad, scare, boyfriend
C9	therapi, depress, therapist, issu, anxieti, disord, eat, month, cost, coupl
C2	anger, forgiv, discuss, hurt, angri, behavior, intak, lie, said, sexual
C19	die, experienc, current, attack, medic, alcohol, rate, daili, health, panic

^aMost frequent and exclusive words that distinguish each topic in patients' transcripts.

Figure 3. Structural topic model estimates of association between psychiatric symptoms and mean topic prevalence. Topics on the right side of the dotted line have higher prevalence in Controls and Health care Workers with moderate to severe anxiety and depression.



Topics and Clinical Levels of Depression and Anxiety

After determining the distribution of topics emerging from treatment transcripts, we examined the association of topics discussed in psychotherapy with patients' moderate-to-severe anxiety or depression at 3 weeks of treatment. Effect estimates and their 95% CIs are reported in [Figure 3](#). Discussion of the hospital and its locations (H5: *unit, hospit, icu*) was significantly more prevalent among HCWs with moderate to severe anxiety or depression (topic prevalence=0.035, 95% CI 0.022-0.048; $P<.001$). This effect was not observed in the matched controls for the pandemic-relevant topic (C25, *pandem, concern, anxiety*: prevalence=0.003, 95% CI -0.012 to 0.018; $P=.67$), nor for the work-relevant topic (C27, *team, manag, boss*: prevalence=-0.005, 95% CI -0.021 to 0.011; $P=.55$). Other topics were significantly more prevalent for symptomatic HCWs, including endorsing affective disturbances (H15, *depress, feel, mood*: prevalence=0.014, 95% CI 0.002-0.026; $P=.03$) and sleep disturbances (H13, *sleep, night, bed*: prevalence=0.016, 95% CI 0.002-0.030; $P=.02$). No other mental health and health care topics occurred at significantly higher frequency in HCWs with moderate to severe anxiety or depression, with weak trends observed for discussions related to panic attacks, grief, and mask-related concerns. Controls with moderate to severe symptoms were more likely to discuss affective disturbances (C16: prevalence=0.021, 95% CI 0.005-0.037; $P=.01$). Numeric estimates for all HCWs and control topics are reported in [Multimedia Appendix 1](#) [5,29,36-49].

Discussion

Overview

In this study, we examined topics for 820 HCWs and 820 matched general-population outpatients undergoing psychotherapy through a telehealth platform in spring 2020 during the first US wave of COVID-19 and their associations with moderate to severe depression and anxiety. In total, 3 weeks of treatment transcripts were examined using NLP methods, enabling elucidation of the content of therapy discussions automatically at scale and in a privacy-preserving way. Results indicated significant differences in the proportion of health care-related topics between HCW and control cohorts, as well as their association with moderate to severe anxiety and depression.

Analysis of the distribution of NLP-derived treatment topics indicated that HCWs extensively discussed health care-related topics in psychotherapy. Specifically, HCWs had 4 conversational themes around health care, while controls only had 1. This finding is consistent with the increased work-related stressors experienced by HCWs during the COVID-19 pandemic, when they were particularly vulnerable to work-related adverse impacts compared to the general population, given the increased professional and personal responsibilities they faced. These unique effects of COVID-19 made HCWs specifically vulnerable to mental health problems compared to the general population [53]. In addition to the effect of potential stressors experienced by HCWs during the COVID-19 pandemic, this finding is also consistent with prior

literature indicating that work-related stress is almost twice as prevalent for HCWs as for workers in other fields after controlling for work hours, with physicians at the front line of care at greatest risk [54]. Unique factors that contribute to this include longer hours and greater difficulty with work-life integration compared to other US workers.

Analysis of the prevalence of topics and their association with symptomatology indicated that among HCWs, discussion of hospital settings was significantly associated with moderate to severe anxiety and depression. This association was unique for HCWs and not present in the general-population outpatients, despite shared anxiety, work, and health-related concerns during the pandemic [55]. Discussion of sleep disturbances and mood difficulties were also significantly associated with moderate to severe anxiety and depression in HCWs. These findings confirm the connection between anxiety, depression, and concerns related to being a practicing medical professional during the COVID-19 pandemic [56]. Although not assessed here, possible underlying contributing factors may be hypothesized to include longer exposures to stressful working environments, a higher level of personal responsibility in critical situations, and increased sleep disruption [19]. Sleep deprivation among HCWs has been consistently linked to increases in anxiety, depression, and suicidal ideation [57]. These findings are especially robust for HCWs who work longer hours, who work night shifts, and who have less time off between their shifts [58]. Existing literature supports a similar relationship for how work-related stress and anxiety and depressive symptoms mutually reinforce each other [59].

Strengths and Limitations

Findings of this study are unique due to the large corpus of treatment transcripts from HCWs during the initial phase of the COVID-19 pandemic, and data analytic methods exploring the use of computational linguistics to identify stated risk factors. To the growing body of literature documenting the challenges posed to mental health and well-being by the COVID pandemic, we contribute a proof-of-concept demonstrating that web-based therapy platforms can serve as unique observatories for the mental health needs of hard-to-reach populations like HCWs. This study has several limitations. First, our sample consisted of self-referred patients, and differences in access to telehealth services could reduce the generalizability of results. Second, our sample showed a skew toward female individuals and nursing occupations, although this distribution aligned consistently with US population occupational statistics for HCWs [60]. Third, we focused our analysis on a concatenation of all of a patient's talk turns during the first 3 weeks of treatment. Future work should focus on complex modeling of topics over time, for example using sequential models to examine topics turn-by-turn, as well as models incorporating therapists' talk turns. Fourth, our findings emerged from the corpus the STM was trained on and might not generalize when applied to different corpora, such as transcripts in languages other than English. Future studies should consider using pretrained large language models on wider corpora of clinical data for more generalizable topic representations across multiple domains and languages [61]. Fifth, topic associations with symptoms were limited to data from validated self-report

measures, and other methods to capture psychiatric symptoms may return different results.

Privacy and Ethics

Important ethical considerations about patient privacy need to be made when accessing sensitive health information such as psychotherapy transcripts. This study included several privacy-preserving measures to reduce risks associated with the study. First, all patients and clinicians gave informed consent to the use of their data in a deidentified and aggregated format for research purposes as part of the user agreement they signed before they began using the platform. All procedures were approved by the university institutional review board. Second, all transcripts were deidentified by the platform prior to the research team accessing the data. Deidentification removed any personal identifiers, like proper nouns, locations, and dates, among other potential identifiers. Third, we limited our analyses to the outputs of STM, which are distributions of common words less likely to reveal private information than the raw text. The first 2 authors (MM and ET) handled the primary analyses and were the only authors to view any portion of raw deidentified text, accessed exclusively as part of model development. Fourth, HCWs' NPIs and associated information were not accessed as part of the study. Rather, specific health care occupations were identified using named entity recognition on the deidentified transcripts. This solution allowed us to extract occupational information while minimizing access to the raw deidentified transcripts, thus further preserving patient privacy.

Conclusions

Among US HCWs seeking psychotherapy treatment in spring 2020 during the first wave of the COVID-19 pandemic, discussion of workplace-related concerns was uniquely associated with moderate to severe anxiety and depression. The association between health care work and psychiatric symptoms was unique, going beyond other quality-of-life factors potentially related to work such as poor sleep hygiene. We contribute to the literature on the psychological burden associated with health care work by demonstrating that HCW-specific content related to anxiety and depression emerges naturally in the context of web-based psychotherapy. These findings highlight the unique mental health concerns faced by HCWs during the COVID-19 pandemic, a time with significantly increased work demands, lack of social support, and fear of infection from work activity for HCWs and their families. These stressors were in addition to work-related stressors regularly faced by HCWs [54]. The results of this research could help pinpoint the key factors contributing to the

high levels of depression and anxiety among HCWs and fill the gaps in care. The increased stress put on HCWs during COVID-19 along with the established link between HCWs' mental health and societal well-being supports the critical need to prioritize mental health treatment provision for HCWs.

As mental health risk factors were captured automatically from transcripts using NLP methods, the study also serves as a proof of concept for the automated detection of psychological distress in HCWs. One of the main advantages of NLP markers is that they can identify specific language patterns that are associated with anxiety and depression. Unlike traditional assessment methods, such as self-reported surveys and interviews, NLP markers from psychotherapy platforms present a passive and less burdensome way to assess therapy-seekers' mental health, akin to the digital biomarkers of mental health researchers have developed from wearable and smartphone data. Defining and validating NLP markers of anxiety and depression could lead to more accurate and reliable assessments, which would be beneficial for both patients and health care providers. Moreover, NLP markers could help to better understand the underlying mechanisms of anxiety and depression by teasing patients into different subgroups based on their specific needs and characteristics. By identifying these patterns, we could tailor treatment and intervention strategies to the specific needs of each patient in clinical settings [62,63]. Eventually, NLP methods could support the advancement of personalized medicine approaches where mental health needs can be estimated routinely using automated methods in ecological or real-world settings. This could be achieved by designing digital apps [64] that offer periodic checks to elicit narrative content about potential risk factors and stressors. Transcripts of the narratives could then be analyzed to extract conversational topics associated with probabilities of experiencing distress through NLP techniques such as STM. For example, this approach could identify language patterns focusing on work-related stressors (such as our HCW sample) or behavioral disturbances (eg, poor sleep hygiene), and then offer personalized triage and resource recommendations. Similar work has been conducted in the context of crisis counseling platforms to analyze patients' messages for suicidal ideation [65,66]. Offering mental health resources at scale through automated recommendations could help HCWs overcome barriers to treatment access including stigma and unpredictable work hours. Given the high-stress nature of the health care profession, there is vast potential for designing automated systems that can proactively evaluate individual needs and provide personalized resources for preventive care.

Acknowledgments

MM's research was supported by the National Center for Advancing Translational Sciences and the National Institutes of Health (grants 2KL2TR001446-06A1 and 1K23MH134068-01), Talkspace, and by the American Foundation for Suicide Prevention (grant PRG-0-104-19). ET is supported by a Microsoft Research PhD Fellowship and a Digital Life Initiative Doctoral Fellowship. TDH's research was supported by National Institutes of Health (awards R44MH124334 and R01MH125179-01). TKC is a cofounder and equity holder of HealthRhythms, Inc, is coemployed by UnitedHealth Group, and has received grants from Click Therapeutics related to digital therapeutics outside the submitted work. NMS reports research support from the Department of Defense, the Patient-Centered Outcomes Research Institute, and the National Institutes of Health; in addition, she reports consulting for Axovant Sciences, Springworks, Praxis Therapeutics, Aptinix, Genomind, Wolters Kluwer (royalty), and spousal equity in

G1 Therapeutics. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

Data Availability

Deidentified patient data may be made available upon completion of a data use agreement and data security review with Talkspace. Analytic code describing natural language processing methods and algorithms is available in [Multimedia Appendices 2](#) and [3](#).

Authors' Contributions

All authors contributed to the study concept and design. MM, TKC, and NMS supervised the study. MM and TDH acquired the data. ET and MM analyzed and interpreted the data, and take responsibility for the integrity of the data and the accuracy of the data analyses. TDH provided administrative, technical, and material support. All authors contributed to the drafting of the paper and its critical revision for important intellectual content.

Conflicts of Interest

TDH is an employee of the platform that provided the data. Talkspace had no role in the analysis, interpretation of the data, or decision to submit the paper for publication. TKC is a cofounder and equity holder of HealthRhythms, Inc.

Multimedia Appendix 1

Study sample flow chart and details on structural topic modeling.

[[PDF File \(Adobe PDF File\), 1405 KB - ai_v2i1e47223_app1.pdf](#)]

Multimedia Appendix 2

Analytic code: topic modeling.

[[PDF File \(Adobe PDF File\), 1756 KB - ai_v2i1e47223_app2.pdf](#)]

Multimedia Appendix 3

Analytic code: job identification algorithm.

[[PDF File \(Adobe PDF File\), 1217 KB - ai_v2i1e47223_app3.pdf](#)]

References

1. WHO coronavirus COVID-19 dashboard. World Health Organization. 2021. URL: <https://covid19.who.int/> [accessed 2022-06-21]
2. Grimm CA. Hospitals reported that the COVID-19 pandemic has significantly strained health care delivery. Office of Inspector General. 2021. URL: <https://oig.hhs.gov/oei/reports/OEI-09-21-00140.asp> [accessed 2023-09-19]
3. Hospital utilization. HHS Protect Public Data Hub. 2020. URL: <https://protect-public.hhs.gov/pages/hospital-utilization> [accessed 2021-10-14]
4. Lai X, Wang M, Qin C, Tan L, Ran L, Chen D, et al. Coronavirus disease 2019 (COVID-2019) infection among health care workers and implications for prevention measures in a tertiary hospital in Wuhan, China. *JAMA Netw Open* 2020;3(5):e209666 [FREE Full text] [doi: [10.1001/jamanetworkopen.2020.9666](https://doi.org/10.1001/jamanetworkopen.2020.9666)] [Medline: [32437575](https://pubmed.ncbi.nlm.nih.gov/32437575/)]
5. Luo M, Guo L, Yu M, Jiang W, Wang H. The psychological and mental impact of coronavirus disease 2019 (COVID-19) on medical staff and general public - a systematic review and meta-analysis. *Psychiatry Res* 2020;291:113190 [FREE Full text] [doi: [10.1016/j.psychres.2020.113190](https://doi.org/10.1016/j.psychres.2020.113190)] [Medline: [32563745](https://pubmed.ncbi.nlm.nih.gov/32563745/)]
6. Pappa S, Ntella V, Giannakas T, Giannakoulis VG, Papoutsis E, Katsaounou P. Prevalence of depression, anxiety, and insomnia among healthcare workers during the COVID-19 pandemic: a systematic review and meta-analysis. *Brain Behav Immun* 2020;88:901-907 [FREE Full text] [doi: [10.1016/j.bbi.2020.05.026](https://doi.org/10.1016/j.bbi.2020.05.026)] [Medline: [32437915](https://pubmed.ncbi.nlm.nih.gov/32437915/)]
7. Raudenská J, Steinerová V, Javůrková A, Urits I, Kaye AD, Viswanath O, et al. Occupational burnout syndrome and post-traumatic stress among healthcare professionals during the novel coronavirus disease 2019 (COVID-19) pandemic. *Best Pract Res Clin Anaesthesiol* 2020;34(3):553-560 [FREE Full text] [doi: [10.1016/j.bpa.2020.07.008](https://doi.org/10.1016/j.bpa.2020.07.008)] [Medline: [33004166](https://pubmed.ncbi.nlm.nih.gov/33004166/)]
8. Sterling MR, Tseng E, Poon A, Cho J, Avgar AC, Kern LM, et al. Experiences of home health care workers in New York City during the coronavirus disease 2019 pandemic: a qualitative analysis. *JAMA Intern Med* 2020;180(11):1453-1459 [FREE Full text] [doi: [10.1001/jamainternmed.2020.3930](https://doi.org/10.1001/jamainternmed.2020.3930)] [Medline: [32749450](https://pubmed.ncbi.nlm.nih.gov/32749450/)]
9. de Pablo GS, Vaquerizo-Serrano J, Catalan A, Arango C, Moreno C, Ferre F, et al. Impact of coronavirus syndromes on physical and mental health of health care workers: systematic review and meta-analysis. *J Affect Disord* 2020;275:48-57 [FREE Full text] [doi: [10.1016/j.jad.2020.06.022](https://doi.org/10.1016/j.jad.2020.06.022)] [Medline: [32658823](https://pubmed.ncbi.nlm.nih.gov/32658823/)]
10. Moazzami B, Razavi-Khorasani N, Moghadam AD, Farokhi E, Rezaei N. COVID-19 and telemedicine: immediate action required for maintaining healthcare providers well-being. *J Clin Virol* 2020;126:104345 [FREE Full text] [doi: [10.1016/j.jcv.2020.104345](https://doi.org/10.1016/j.jcv.2020.104345)] [Medline: [32278298](https://pubmed.ncbi.nlm.nih.gov/32278298/)]

11. Patel RS, Bachu R, Adikey A, Malik M, Shah M. Factors related to physician burnout and its consequences: a review. *Behav Sci (Basel)* 2018;8(11):98 [FREE Full text] [doi: [10.3390/bs8110098](https://doi.org/10.3390/bs8110098)] [Medline: [30366419](https://pubmed.ncbi.nlm.nih.gov/30366419/)]
12. Wallace JE, Lemaire JB, Ghali WA. Physician wellness: a missing quality indicator. *Lancet* 2009;374(9702):1714-1721 [FREE Full text] [doi: [10.1016/S0140-6736\(09\)61424-0](https://doi.org/10.1016/S0140-6736(09)61424-0)] [Medline: [19914516](https://pubmed.ncbi.nlm.nih.gov/19914516/)]
13. Mata DA, Ramos MA, Bansal N, Khan R, Guille C, Di Angelantonio E, et al. Prevalence of depression and depressive symptoms among resident physicians: a systematic review and meta-analysis. *JAMA* 2015;314(22):2373-2383 [FREE Full text] [doi: [10.1001/jama.2015.15845](https://doi.org/10.1001/jama.2015.15845)] [Medline: [26647259](https://pubmed.ncbi.nlm.nih.gov/26647259/)]
14. Dutheil F, Aubert C, Pereira B, Dambun M, Moustafa F, Mermillod M, et al. Suicide among physicians and health-care workers: a systematic review and meta-analysis. *PLoS One* 2019;14(12):e0226361 [FREE Full text] [doi: [10.1371/journal.pone.0226361](https://doi.org/10.1371/journal.pone.0226361)] [Medline: [31830138](https://pubmed.ncbi.nlm.nih.gov/31830138/)]
15. Brand SL, Coon JT, Fleming LE, Carroll L, Bethel A, Wyatt K. Whole-system approaches to improving the health and wellbeing of healthcare workers: a systematic review. *PLoS One* 2017;12(12):e0188418 [FREE Full text] [doi: [10.1371/journal.pone.0188418](https://doi.org/10.1371/journal.pone.0188418)] [Medline: [29200422](https://pubmed.ncbi.nlm.nih.gov/29200422/)]
16. Zaçe D, Hoxhaj I, Orfino A, Viteritti AM, Janiri L, Di Pietro ML. Interventions to address mental health issues in healthcare workers during infectious disease outbreaks: a systematic review. *J Psychiatr Res* 2021;136:319-333 [FREE Full text] [doi: [10.1016/j.jpsychires.2021.02.019](https://doi.org/10.1016/j.jpsychires.2021.02.019)] [Medline: [33636688](https://pubmed.ncbi.nlm.nih.gov/33636688/)]
17. Fahrenkopf AM, Sectish TC, Barger LK, Sharek PJ, Lewin D, Chiang VW, et al. Rates of medication errors among depressed and burnt out residents: prospective cohort study. *BMJ* 2008;336(7642):488-491 [FREE Full text] [doi: [10.1136/bmj.39469.763218.BE](https://doi.org/10.1136/bmj.39469.763218.BE)] [Medline: [18258931](https://pubmed.ncbi.nlm.nih.gov/18258931/)]
18. Hall LH, Johnson J, Watt I, Tsipa A, O'Connor DB. Healthcare staff wellbeing, burnout, and patient safety: a systematic review. *PLoS One* 2016;11(7):e0159015 [FREE Full text] [doi: [10.1371/journal.pone.0159015](https://doi.org/10.1371/journal.pone.0159015)] [Medline: [27391946](https://pubmed.ncbi.nlm.nih.gov/27391946/)]
19. Weaver MD, Vetter C, Rajaratnam SMW, O'Brien CS, Qadri S, Benca RM, et al. Sleep disorders, depression and anxiety are associated with adverse safety outcomes in healthcare workers: a prospective cohort study. *J Sleep Res* 2018;27(6):e12722 [FREE Full text] [doi: [10.1111/jsr.12722](https://doi.org/10.1111/jsr.12722)] [Medline: [30069960](https://pubmed.ncbi.nlm.nih.gov/30069960/)]
20. West CP, Tan AD, Habermann TM, Sloan JA, Shanafelt TD. Association of resident fatigue and distress with perceived medical errors. *JAMA* 2009;302(12):1294-1300 [FREE Full text] [doi: [10.1001/jama.2009.1389](https://doi.org/10.1001/jama.2009.1389)] [Medline: [19773564](https://pubmed.ncbi.nlm.nih.gov/19773564/)]
21. Center C, Davis M, Detre T, Ford DE, Hansbrough W, Hendin H, et al. Confronting depression and suicide in physicians: a consensus statement. *JAMA* 2003;289(23):3161-3166 [FREE Full text] [doi: [10.1001/jama.289.23.3161](https://doi.org/10.1001/jama.289.23.3161)] [Medline: [12813122](https://pubmed.ncbi.nlm.nih.gov/12813122/)]
22. Miles SH. A piece of my mind. a challenge to licensing boards: the stigma of mental illness. *JAMA* 1998;280(10):865 [FREE Full text] [doi: [10.1001/jama.280.10.865](https://doi.org/10.1001/jama.280.10.865)] [Medline: [9739951](https://pubmed.ncbi.nlm.nih.gov/9739951/)]
23. Wimsatt LA, Schwenk TL, Sen A. Predictors of depression stigma in medical students: potential targets for prevention and education. *Am J Prev Med* 2015;49(5):703-714 [FREE Full text] [doi: [10.1016/j.amepre.2015.03.021](https://doi.org/10.1016/j.amepre.2015.03.021)] [Medline: [26141915](https://pubmed.ncbi.nlm.nih.gov/26141915/)]
24. Torous J, Bucci S, Bell IH, Kessing LV, Faurholt-Jepsen M, Whelan P, et al. The growing field of digital psychiatry: current evidence and the future of apps, social media, chatbots, and virtual reality. *World Psychiatry* 2021;20(3):318-335 [FREE Full text] [doi: [10.1002/wps.20883](https://doi.org/10.1002/wps.20883)] [Medline: [34505369](https://pubmed.ncbi.nlm.nih.gov/34505369/)]
25. Cho CH, Lee T, Kim MG, In HP, Kim L, Lee HJ. Mood prediction of patients with mood disorders by machine learning using passive digital phenotypes based on the circadian rhythm: prospective observational cohort study. *J Med Internet Res* 2019;21(4):e11029 [FREE Full text] [doi: [10.2196/11029](https://doi.org/10.2196/11029)] [Medline: [30994461](https://pubmed.ncbi.nlm.nih.gov/30994461/)]
26. Ren B, Xia CH, Gehrman P, Barnett I, Satterthwaite T. Measuring daily activity rhythms in young adults at risk of affective instability using passively collected smartphone data: observational study. *JMIR Form Res* 2022;6(9):e33890 [FREE Full text] [doi: [10.2196/33890](https://doi.org/10.2196/33890)] [Medline: [36103225](https://pubmed.ncbi.nlm.nih.gov/36103225/)]
27. Adler DA, Tseng VWS, Qi G, Scarpa J, Sen S, Choudhury T. Identifying mobile sensing indicators of stress-resilience. *Proc ACM Interact Mob Wearable Ubiquitous Technol* 2021;5(2):1-32 [FREE Full text] [doi: [10.1145/3463528](https://doi.org/10.1145/3463528)] [Medline: [35445162](https://pubmed.ncbi.nlm.nih.gov/35445162/)]
28. Malgaroli M, Hull TD, Zech JM, Althoff T. Natural language processing for mental health interventions: a systematic review and research framework. *Transl Psychiatry* 2023 Oct 06;13(1):309 [FREE Full text] [doi: [10.1038/s41398-023-02592-2](https://doi.org/10.1038/s41398-023-02592-2)] [Medline: [37798296](https://pubmed.ncbi.nlm.nih.gov/37798296/)]
29. Blei DM. Probabilistic topic models. *Commun ACM* 2012;55(4):77-84 [FREE Full text] [doi: [10.1145/2133806.2133826](https://doi.org/10.1145/2133806.2133826)]
30. Jelodar H, Wang Y, Yuan C, Feng X, Jiang X, Li Y, et al. Latent Dirichlet Allocation (LDA) and topic modeling: models, applications, a survey. *Multimed Tools Appl* 2019;78(11):15169-15211. [doi: [10.1007/s11042-018-6894-4](https://doi.org/10.1007/s11042-018-6894-4)]
31. Gong Y, Poellabauer C. Topic modeling based multi-modal depression detection. 2017 Presented at: MM '17: ACM Multimedia Conference; 23 October 2017; Mountain View California USA p. 69-76. [doi: [10.1145/3133944.3133945](https://doi.org/10.1145/3133944.3133945)]
32. Shen JH, Rudzicz F. Detecting anxiety through reddit. 2017 Presented at: Proceedings of the Fourth Workshop on Computational Linguistics and Clinical Psychology — From Linguistic Signal to Clinical Reality; August 2017; Vancouver, BC p. 58-65. [doi: [10.18653/v1/w17-3107](https://doi.org/10.18653/v1/w17-3107)]
33. Low DM, Rumker L, Talkar T, Torous J, Cecchi G, Ghosh SS. Natural language processing reveals vulnerable mental health support groups and heightened health anxiety on reddit during COVID-19: observational study. *J Med Internet Res* 2020;22(10):e22635 [FREE Full text] [doi: [10.2196/22635](https://doi.org/10.2196/22635)] [Medline: [32936777](https://pubmed.ncbi.nlm.nih.gov/32936777/)]

34. Insel TR. Digital phenotyping: a global tool for psychiatry. *World Psychiatry* 2018;17(3):276-277 [FREE Full text] [doi: [10.1002/wps.20550](https://doi.org/10.1002/wps.20550)] [Medline: [30192103](https://pubmed.ncbi.nlm.nih.gov/30192103/)]
35. Talkspace. URL: <https://www.talkspace.com/> [accessed 2023-09-22]
36. Posner K, Brown GK, Stanley B, Brent DA, Yershova KV, Oquendo MA, et al. The Columbia-suicide severity rating scale: initial validity and internal consistency findings from three multisite studies with adolescents and adults. *Am J Psychiatry* 2011;168(12):1266-1277 [FREE Full text] [doi: [10.1176/appi.ajp.2011.10111704](https://doi.org/10.1176/appi.ajp.2011.10111704)] [Medline: [22193671](https://pubmed.ncbi.nlm.nih.gov/22193671/)]
37. R Core Team. R: A Language and Environment for Statistical Computing. URL: <https://www.R-project.org/> [accessed 2023-09-19]
38. Roberts ME, Stewart BM, Tingley D. Stm: an R package for structural topic models. *J Stat Softw* 2019;91:1-40 [FREE Full text] [doi: [10.18637/jss.v091.i02](https://doi.org/10.18637/jss.v091.i02)]
39. Airoldi EM, Bischof JM. Improving and evaluating topic models and other models of text. *J Am Stat Assoc* 2016;111(516):1381-1403 [FREE Full text] [doi: [10.1080/01621459.2015.1051182](https://doi.org/10.1080/01621459.2015.1051182)]
40. Chang J, Boyd-Graber J, Gerrish S, Wang C, Blei DM. Reading tea leaves: how humans interpret topic models. 2009 Presented at: 22nd International Conference on Neural Information Processing Systems; December 7, 2009; Vancouver, BC.
41. Grimmer, Stewart. Text as data: The promise and pitfalls of automatic content analysis methods for political texts. *Polit Anal* 2013;21(3):267-297 [FREE Full text]
42. NLP applied to occupational health: MEDDOPROF shared task at IberLEF 2021 on automatic recognition, classification and normalization of professions and occupations from medical texts. *Sociedad Española para el Procesamiento del Lenguaje Natural*. URL: <http://journal.sepln.org/sepln/ojs/ojs/index.php/pln/article/view/6393> [accessed 2023-10-04]
43. Mimno D, Wallach H, Talley E, Leenders M, McCallum A. Optimizing semantic coherence in topic models. 2011 Presented at: 2011 Conference on Empirical Methods in Natural Language Processing; July 2011; Edinburgh, Scotland, UK URL: <https://aclanthology.org/D11-1024.pdf>
44. Murgado AM, Portillo AP, Úbeda PL, Martín M, Ureña-López A. Identifying professions and occupations in health-related social media using natural language processing. 2021 Presented at: Sixth Social Media Mining for Health (#SMM4H) Workshop and Shared Task; June 2021; Mexico City, Mexico URL: <https://aclanthology.org/2021.smm4h-1.31.pdf>
45. Roberts M, Stewart B, Tingley D, Airoldi E. The structural topic model and applied social science. *Neural Information Processing Society*. URL: <https://scholar.harvard.edu/sites/scholar.harvard.edu/files/dtingley/files/stmnips2013.pdf> [accessed 2023-10-04]
46. Schoene AM, Basinas I, van Tongeren M, Ananiadou S. A narrative literature review of natural language processing applied to the occupational exposome. *Int J Environ Res Public Health* 2022 Jul 13;19(14):8544 [FREE Full text] [doi: [10.3390/ijerph19148544](https://doi.org/10.3390/ijerph19148544)] [Medline: [35886395](https://pubmed.ncbi.nlm.nih.gov/35886395/)]
47. Ho D, Imai K, King G, Stuart EA. Matchit: Nonparametric preprocessing for parametric causal inference. *J Stat Softw* 2011;42(8):1-28 [FREE Full text] [doi: [10.18637/JSS.V042.I08](https://doi.org/10.18637/JSS.V042.I08)]
48. Talkspace is donating free therapy to medical workers fighting COVID-19. *Talkspace*. URL: <https://www.talkspace.com/blog/coronavirus-talkspace-donation-healthcare-workers/> [accessed 2023-10-04]
49. Thoemmes FJ, Kim ES. A systematic review of propensity score methods in the social sciences. *Multivariate Behav Res* 2011 Feb 07;46(1):90-118. [doi: [10.1080/00273171.2011.540475](https://doi.org/10.1080/00273171.2011.540475)] [Medline: [26771582](https://pubmed.ncbi.nlm.nih.gov/26771582/)]
50. Kroenke K, Spitzer RL, Williams JB. The PHQ-9: validity of a brief depression severity measure. *J Gen Intern Med* 2001;16(9):606-613 [FREE Full text] [doi: [10.1046/j.1525-1497.2001.016009606.x](https://doi.org/10.1046/j.1525-1497.2001.016009606.x)] [Medline: [11556941](https://pubmed.ncbi.nlm.nih.gov/11556941/)]
51. Spitzer RL, Kroenke K, Williams JBW, Löwe B. A brief measure for assessing generalized anxiety disorder: the GAD-7. *Arch Intern Med* 2006;166(10):1092-1097 [FREE Full text] [doi: [10.1001/archinte.166.10.1092](https://doi.org/10.1001/archinte.166.10.1092)] [Medline: [16717171](https://pubmed.ncbi.nlm.nih.gov/16717171/)]
52. Caspi A, Moffitt TE. All for one and one for all: mental disorders in one dimension. *Am J Psychiatry* 2018;175(9):831-844 [FREE Full text] [doi: [10.1176/appi.ajp.2018.17121383](https://doi.org/10.1176/appi.ajp.2018.17121383)] [Medline: [29621902](https://pubmed.ncbi.nlm.nih.gov/29621902/)]
53. Krishnamoorthy Y, Nagarajan R, Saya GK, Menon V. Prevalence of psychological morbidities among general population, healthcare workers and COVID-19 patients amidst the COVID-19 pandemic: a systematic review and meta-analysis. *Psychiatry Res* 2020;293:113382 [FREE Full text] [doi: [10.1016/j.psychres.2020.113382](https://doi.org/10.1016/j.psychres.2020.113382)] [Medline: [32829073](https://pubmed.ncbi.nlm.nih.gov/32829073/)]
54. Shanafelt TD, Boone S, Tan L, Dyrbye LN, Sotile W, Satele D, et al. Burnout and satisfaction with work-life balance among US physicians relative to the general US population. *Arch Intern Med* 2012;172(18):1377-1385 [FREE Full text] [doi: [10.1001/archinternmed.2012.3199](https://doi.org/10.1001/archinternmed.2012.3199)] [Medline: [22911330](https://pubmed.ncbi.nlm.nih.gov/22911330/)]
55. Hull TD, Levine J, Bantilan N, Desai AN, Majumder MS. Analyzing digital evidence from a telemental health platform to assess complex psychological responses to the COVID-19 pandemic: content analysis of text messages. *JMIR Form Res* 2021;5(2):e26190 [FREE Full text] [doi: [10.2196/26190](https://doi.org/10.2196/26190)] [Medline: [33502999](https://pubmed.ncbi.nlm.nih.gov/33502999/)]
56. Marvaldi M, Mallet J, Dubertret C, Moro MR, Guessoum SB. Anxiety, depression, trauma-related, and sleep disorders among healthcare workers during the COVID-19 pandemic: a systematic review and meta-analysis. *Neurosci Biobehav Rev* 2021;126:252-264 [FREE Full text] [doi: [10.1016/j.neubiorev.2021.03.024](https://doi.org/10.1016/j.neubiorev.2021.03.024)] [Medline: [33774085](https://pubmed.ncbi.nlm.nih.gov/33774085/)]
57. Booker LA, Magee M, Rajaratnam SMW, Sletten TL, Howard ME. Individual vulnerability to insomnia, excessive sleepiness and shift work disorder amongst healthcare shift workers. a systematic review. *Sleep Med Rev* 2018;41:220-233 [FREE Full text] [doi: [10.1016/j.smrv.2018.03.005](https://doi.org/10.1016/j.smrv.2018.03.005)] [Medline: [29680177](https://pubmed.ncbi.nlm.nih.gov/29680177/)]

58. Eldevik MF, Flo E, Moen BE, Pallesen S, Bjorvatn B. Insomnia, excessive sleepiness, excessive fatigue, anxiety, depression and shift work disorder in nurses having less than 11 hours in-between shifts. *PLoS One* 2013;8(8):e70882 [FREE Full text] [doi: [10.1371/journal.pone.0070882](https://doi.org/10.1371/journal.pone.0070882)] [Medline: [23976964](https://pubmed.ncbi.nlm.nih.gov/23976964/)]
59. Bianchi R, Schonfeld IS, Laurent E. Burnout-depression overlap: a review. *Clin Psychol Rev* 2015;36:28-41 [FREE Full text] [doi: [10.1016/j.cpr.2015.01.004](https://doi.org/10.1016/j.cpr.2015.01.004)] [Medline: [25638755](https://pubmed.ncbi.nlm.nih.gov/25638755/)]
60. Laughlin L, Anderson A, Martinez A, Gayfield A. 22 Million employed in health care fight against COVID-19. *Census.gov*. 2021. URL: <https://www.census.gov/library/stories/2021/04/who-are-our-health-care-workers.html> [accessed 2023-06-09]
61. Peinelt N, Nguyen D, Liakata M. tBERT: Topic models and BERT joining forces for semantic similarity detection. 2020 Presented at: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics; July 2020; Online p. 7047-7055. [doi: [10.18653/v1/2020.acl-main.630](https://doi.org/10.18653/v1/2020.acl-main.630)]
62. Sharma A, Lin IW, Miner AS, Atkins DC, Althoff T. Human-AI collaboration enables more empathic conversations in text-based peer-to-peer mental health support. *Nat Mach Intell* 2023;5(1):46-57. [doi: [10.1038/s42256-022-00593-2](https://doi.org/10.1038/s42256-022-00593-2)]
63. Zhang J, Mullainathan S, Danescu-Niculescu-Mizil C. Quantifying the causal effects of conversational tendencies. *Proc ACM Hum-Comput Interact* 2020;4(CSCW2):1-24 [FREE Full text] [doi: [10.1145/3415202](https://doi.org/10.1145/3415202)]
64. Aung MH, Matthews M, Choudhury T. Sensing behavioral symptoms of mental health and delivering personalized interventions using mobile technologies. *Depress Anxiety* 2017;34(7):603-609 [FREE Full text] [doi: [10.1002/da.22646](https://doi.org/10.1002/da.22646)] [Medline: [28661072](https://pubmed.ncbi.nlm.nih.gov/28661072/)]
65. Althoff T, Clark K, Leskovec J. Large-scale analysis of counseling conversations: an application of natural language processing to mental health. *Trans Assoc Comput Linguist* 2016;4:463-476 [FREE Full text] [Medline: [28344978](https://pubmed.ncbi.nlm.nih.gov/28344978/)]
66. Bantilan N, Malgaroli M, Ray B, Hull TD. Just in time crisis response: suicide alert system for telemedicine psychotherapy settings. *Psychother Res* 2020 Jun 19;31(3):289-299. [doi: [10.1080/10503307.2020.1781952](https://doi.org/10.1080/10503307.2020.1781952)] [Medline: [32558625](https://pubmed.ncbi.nlm.nih.gov/32558625/)]

Abbreviations

GAD-7: General Anxiety Disorder Scale-7

HCW: health care worker

ML: machine learning

NLP: natural language processing

NPI: National Provider Identifier

PHQ-9: Patient Health Questionnaire-9

STM: structural topic model

HIPAA: Health Insurance Portability and Accountability Act

Edited by H Liu; submitted 12.03.23; peer-reviewed by K Schultebrasucks, E Korshakova; comments to author 03.06.23; revised version received 28.06.23; accepted 07.09.23; published 24.10.23.

Please cite as:

Malgaroli M, Tseng E, Hull TD, Jennings E, Choudhury TK, Simon NM

Association of Health Care Work With Anxiety and Depression During the COVID-19 Pandemic: Structural Topic Modeling Study

JMIR AI 2023;2:e47223

URL: <https://ai.jmir.org/2023/1/e47223>

doi: [10.2196/47223](https://doi.org/10.2196/47223)

PMID: [38875560](https://pubmed.ncbi.nlm.nih.gov/38875560/)

©Matteo Malgaroli, Emily Tseng, Thomas D Hull, Emma Jennings, Tanzeem K Choudhury, Naomi M Simon. Originally published in JMIR AI (<https://ai.jmir.org>), 24.10.2023. This is an open-access article distributed under the terms of the Creative Commons Attribution License (<https://creativecommons.org/licenses/by/4.0/>), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR AI, is properly cited. The complete bibliographic information, a link to the original publication on <https://www.ai.jmir.org/>, as well as this copyright and license information must be included.

Original Paper

Assessing Elevated Blood Glucose Levels Through Blood Glucose Evaluation and Monitoring Using Machine Learning and Wearable Photoplethysmography Sensors: Algorithm Development and Validation

Bohan Shi^{1,2}, PhD; Satvinder Singh Dhaliwal^{3,4,5}, PhD; Marcus Soo¹, BEng; Cheri Chan⁶, BSc; Jocelin Wong¹, BEng; Natalie W C Lam², BSc; Entong Zhou², GCE; Vivien Paitimusa², GCE; Kum Yin Loke², BSc; Joel Chin², BEng; Mei Tuan Chua⁶, MN; Kathy Chiew Suan Liaw⁶, BSc; Amos W H Lim¹, MSc; Fadil Fatin Insyirah⁶, BSc; Shih-Cheng Yen⁷, PhD; Arthur Tay⁸, PhD; Seng Bin Ang^{9,10}, MD

¹Actxa Pte Ltd, Singapore, Singapore

²Activate Interactive Pte Ltd, Singapore, Singapore

³Curtin Health Innovation Research Institute, Curtin University, Perth, Australia

⁴Faculty of Health Sciences, Curtin University, Perth, Australia

⁵Duke-NUS Graduate Medical School, National University of Singapore, Singapore, Singapore

⁶KK Women's and Children's Hospital, Singapore, Singapore

⁷Innovation and Design Programme, Faculty of Engineering, National University of Singapore, Singapore, Singapore

⁸Department of Electrical and Computer Engineering, National University of Singapore, Singapore, Singapore

⁹Family Medicine Academic Clinical Program, Duke-NUS Medical School, Singapore, Singapore

¹⁰Menopause Unit, KK Women's and Children's Hospital, Singapore, Singapore

Corresponding Author:

Bohan Shi, PhD

Actxa Pte Ltd

#13-06A SingPost Center

10 Eunos Road 8

Singapore, 408600

Singapore

Phone: 65 88115658

Email: bohan.sbh@gmail.com

Abstract

Background: Diabetes mellitus is the most challenging and fastest-growing global public health concern. Approximately 10.5% of the global adult population is affected by diabetes, and almost half of them are undiagnosed. The growing at-risk population exacerbates the shortage of health resources, with an estimated 10.6% and 6.2% of adults worldwide having impaired glucose tolerance and impaired fasting glycemia, respectively. All current diabetes screening methods are invasive and opportunistic and must be conducted in a hospital or laboratory by trained professionals. At-risk participants might remain undetected for years and miss the precious time window for early intervention to prevent or delay the onset of diabetes and its complications.

Objective: We aimed to develop an artificial intelligence solution to recognize elevated blood glucose levels (≥ 7.8 mmol/L) noninvasively and evaluate diabetic risk based on repeated measurements.

Methods: This study was conducted at KK Women's and Children's Hospital in Singapore, and 500 participants were recruited (mean age 38.73, SD 10.61 years; mean BMI 24.4, SD 5.1 kg/m²). The blood glucose levels for most participants were measured before and after consuming 75 g of sugary drinks using both a conventional glucometer (Accu-Chek Performa) and a wrist-worn wearable. The results obtained from the glucometer were used as ground-truth measurements. We performed extensive feature engineering on photoplethysmography (PPG) sensor data and identified features that were sensitive to glucose changes. These selected features were further analyzed using an explainable artificial intelligence approach to understand their contribution to our predictions.

Results: Multiple machine learning models were trained and assessed with 10-fold cross-validation, using participant demographic data and critical features extracted from PPG measurements as predictors. A support vector machine with a radial basis function kernel had the best detection performance, with an average accuracy of 84.7%, a sensitivity of 81.05%, a specificity of 88.3%, a precision of 87.51%, a geometric mean of 84.54%, and *F* score of 84.03%.

Conclusions: Our findings suggest that PPG measurements can be used to identify participants with elevated blood glucose measurements and assist in the screening of participants for diabetes risk.

(*JMIR AI* 2023;2:e48340) doi:[10.2196/48340](https://doi.org/10.2196/48340)

KEYWORDS

diabetes mellitus; explainable artificial intelligence; feature engineering; machine learning; photoplethysmography; wearable sensor

Introduction

Diabetes mellitus (DM) is a chronic and heterogeneous metabolic disorder characterized by the presence of hyperglycemia due to deterioration of insulin secretion, defective insulin action, or both [1,2]. There are 3 main types of DM: type 1 DM (T1DM), type 2 DM (T2DM), and gestational diabetes. T2DM is the most prevalent type of diabetes, affecting over 95% of people with diabetes worldwide [3,4].

The prevalence of DM has been proliferating in recent decades, and it is now the most prominent and fastest-growing global public health challenge [5,6]. Uncontrolled diabetes is associated with an increased risk of complications such as cardiovascular disease, kidney failure, vision loss, nerve damage, and overall mortality [7-9]. On the basis of the latest diabetes prevalence estimate, 10.5% of the global adult population is affected by diabetes, and almost half of them are undiagnosed [10]. The growing at-risk population has further strained scarce health resources. Globally, approximately 10.6% of adults have impaired glucose tolerance (IGT) and 6.2% have impaired fasting glycemia (IFG) [4]. IGT and IFG are reversible transitional conditions between normality and diabetes. These conditions, also known as prediabetes, are characterized by elevated blood glucose levels that are not high enough to be classified as diabetes. However, individuals with IGT or IFG are at increased risk of developing cardiovascular disease, coronary heart disease, stroke, and mortality [11]. One of the challenges with IGT and IFG is that they often do not have any obvious symptoms, which means that they can go undetected and undiagnosed for years. Moreover, a follow-up study conducted in Singapore reported that one-third of these individuals with prediabetes would likely develop T2DM within 8 years without lifestyle changes [12]. A similar study with data from the United Kingdom has also reported that a substantial proportion of individuals with prediabetes could progress to T2DM within 5 years [13]. Therefore, predicting the risk of diabetes in the asymptomatic population is a significant health challenge that must be addressed. Early recognition of prediabetes and undiagnosed T2DM will result in a better health outcome or a more favorable long-term prognosis [14].

Currently, the diagnosis of diabetes and prediabetes is well established. T2DM and prediabetes can be detected using one of four methods: (1) the fasting plasma glucose value, (2) the 2-hour plasma glucose value during a 75 g oral glucose tolerance

test, (3) hemoglobin A_{1c}, and (4) a random plasma glucose test [3]. All these diagnostic screening methods are invasive and opportunistic in nature and must be conducted in a hospital or laboratory by trained professionals. A confirmed diagnosis usually requires repeated testing. As all the tests are single-time point screenings, adults aged >35 years are recommended to undergo regular screening every 3 years. Nevertheless, at-risk individuals hardly comply with this recommendation, especially in developing countries, owing to the cost of diagnostic tests and the scarcity of medical resources [15,16].

Unlike T1DM and gestational diabetes, the development of T2DM and its complications is preventable or controllable. A considerable number of studies have shown that lifestyle and behavioral interventions help patients with diabetes achieve adequate glycemic control [17,18]. Recent evidence also suggests that early lifestyle adjustment will help participants with prediabetes return to normoglycemia and reduce the risk of developing T2DM [19-21]. Frequent diabetes screening identifies individuals with a high risk of T2DM 2.2 years earlier [22], creating a precious time frame and opportunity for taking an early intervention to prevent or delay the onset of diabetes and its complications and improve overall clinical outcomes.

For established individuals with diabetes, constant monitoring of their blood glucose concentration is crucial so that appropriate insulin dosage can be administered in a timely manner to avoid acute and chronic complications and delay disease progression. Conventional blood glucose measurement requires patients to prick their fingers several times a day, which causes the development of massive scarring and loss of sensation at the fingertips over the year [23]. This measurement method is invasive, inconvenient, and expensive, which are the main barriers to the effective self-management of diabetes in the older adult group [24,25]. To improve diabetes outcomes and assist patients in self-managing the disease, continuous glucose monitoring devices have entered the market and are made available for some patients with diabetes. However, most continuous glucose monitoring sensors currently available are still invasive, which measures glucose concentration in the subcutis using an electrochemical needle sensor [26]. Users need to replace the sensor frequently and purchase different components of the system regularly, which will cost from US \$2500 to US \$6000 per year [27,28].

In recent years, the advancement and use of wearable technologies and artificial intelligence (AI) have gradually

changed our daily lives, as many people use wrist-worn wearables daily for fitness and health monitoring [29]. Most consumer wearables have incorporated green light reflection photoplethysmography (PPG) sensors into their products. Wearable technology has the potential to greatly expand the impact of public health initiatives by using a proactive approach to identify abnormal physiological signals, assessing disease risk factors, and helping patients manage chronic conditions and recovery [30-33].

In 2011, Monte-Moreno [34] demonstrated the use of PPG data collected using a pulse oximeter to estimate blood glucose levels. By analyzing the PPG waveform, features such as the respiration frequency, heart rate variability (HRV), and other physiological parameters can be extracted. They are then fed into a random forest model, yielding a prediction accuracy of 87.7% based on the Clark error grid. Rodin et al [35] validated a wearable biosensor developed by Zilberstein et al [36] as an indirect measure of glucometry. The biosensor comprises a PPG sensor and an optically sensitive backglass panel that changes its photochemical characteristics according to the concentrations of specific sweat metabolites. In total, 200 adult participants were recruited, and each participant wore a smartwatch to extract PPG data, while blood samples were collected from the antecubital vein concurrently. The estimation of the blood glucose level was derived using a proprietary algorithm developed by SpectroPhon and compared against a glucose lactate analyzer (YSI 2300). The proposed biosensor was able to detect antepandrial glucose with a mean absolute percentage error of 7.4% and a normalized root mean squared error of 11.56%, while postprandial glucose measurements yielded 7.54% mean absolute percentage error and 9.79% normalized root mean squared error. Zhang et al [37] used a smartphone, taking a video of the index finger covering the flash, to capture the fluctuation in the light absorption associated with the change in blood volume. The resulting red, green, and blue image was then transformed into PPG data. The Gaussian fitting method was applied to model the PPG waveform components, from which 28 time-domain and frequency-domain features were

extracted. A support vector machine (SVM) with a Gaussian kernel was trained with data from 80 participants to classify the user's glucose level as normal, borderline, or warning, with an accuracy of 81.49%, 79.85% sensitivity, 83.19% specificity, and 80.2% *F* score. The study was conducted in a highly controlled environment with limited participants, so the generalizability of these results is subject to certain limitations.

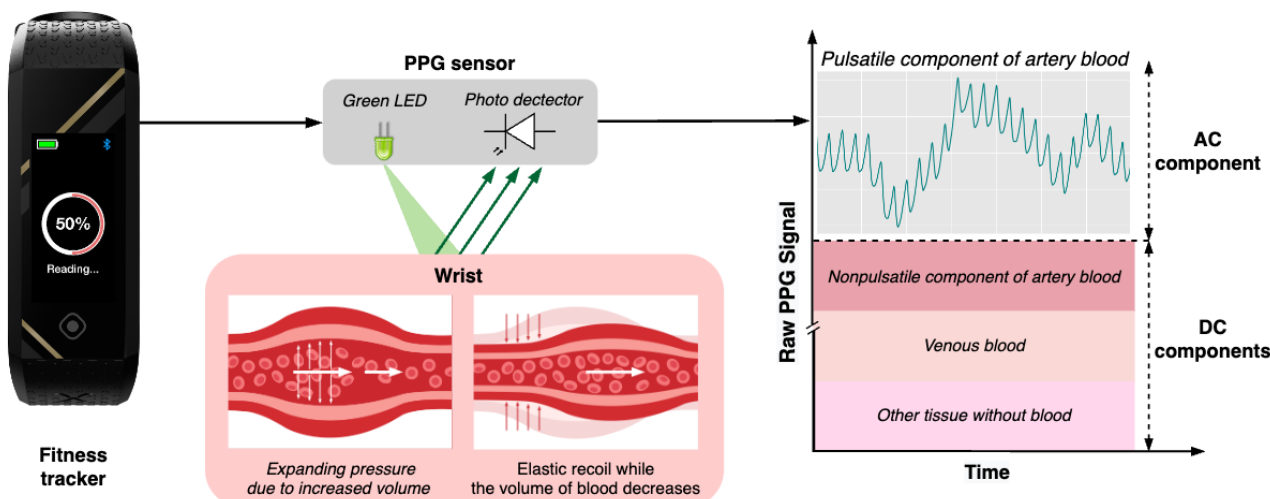
Conventional blood glucose monitoring technologies often require invasive measures such as finger pricking or the use of skin sensors and patches. These methods can be uncomfortable and inconvenient for users and can also be financially burdensome. To address these issues, we propose a novel solution called blood glucose evaluation and monitoring (BGEM) that leverages the latest advancements in signal processing, wearable technology, and AI to detect elevated blood glucose levels and evaluate the risk of developing diabetes. With BGEM, users only need to measure their PPG data using a consumer-grade wrist-worn wearable device. The AI model will then compute relevant digital biomarkers and evaluate the risk of prediabetes or T2DM by recognizing elevated blood glucose levels (≥ 7.8 mmol/L). This solution allows for frequent blood glucose testing without the discomfort and inconvenience of current technologies.

Methods

PPG Sensor

PPG is a low-cost, noninvasive technique that measures the volumetric fluctuation in arterial blood flow [38]. The human wrist is one of the sites for measuring the PPG signal because it has a rich arterial source and an excellent sensor placement with minimal interference to one's daily activities. The PPG signal comprises superimposed pulsatile alternating current components and direct current voltage components. A PPG signal is obtained by illuminating the light emitting device on the skin surface and measuring the variations in light absorption or reflection that reflect the pulsatile flow patterns, as shown in Figure 1.

Figure 1. Illustration of the working principle of a photoplethysmography (PPG) sensor. Changes in blood flow represent different phases within the cardiac cycle. During the diastolic phase, blood volume, arterial diameter, and hemoglobin concentration in the measurement site are minimized, leading to minimum absorption of light by blood and, consequently, an increase in light intensity detected by the sensor system. The reverse is valid for the systolic phase, where a decrease in light intensity is detected instead. AC: alternating current; DC: direct current.



The pulsatile alternating current component corresponds to the cardiac cycle, characterizing that the wrist's blood vessels expand and contract with each heartbeat, whereas the direct current component reflects constant light absorption by venous and arterial blood, as well as other tissues [39]. The PPG signal can detect vascular changes associated with diabetes and contains substantial valuable information from HRV, which is significantly associated with diabetes [40]. Hence, it will be used in this study to extract valuable and meaningful features to identify an individual's glucose status (elevated or normal).

Ethical Considerations

Before commencing the study, ethical clearance was obtained from the SingHealth Centralised Institutional Review Board of Singapore (2020/2968) on March 21, 2021. All methods were performed in accordance with Singapore's clinical guidelines and regulations. Informed consent was obtained from all the trial participants or their legal guardians. The clinical trial was

registered on ClinicalTrials.gov (NCT05504096) on August 17, 2022.

Study Protocol

In total, 500 participants were recruited from KK Women's and Children's Hospital in Singapore. Participants' demographics are summarized in Table 1. For most participants, the blood glucose levels were measured before and after consumption of 75 g of a sugary drink using both the conventional glucometer (Accu-Chek Performa) and the wrist-worn wearable device. Participants who were excluded for the second measurement had high blood glucose measurements ≥ 11.1 mmol/L on their first measurement and hence were not administered the sugary drink measuring 75 g.

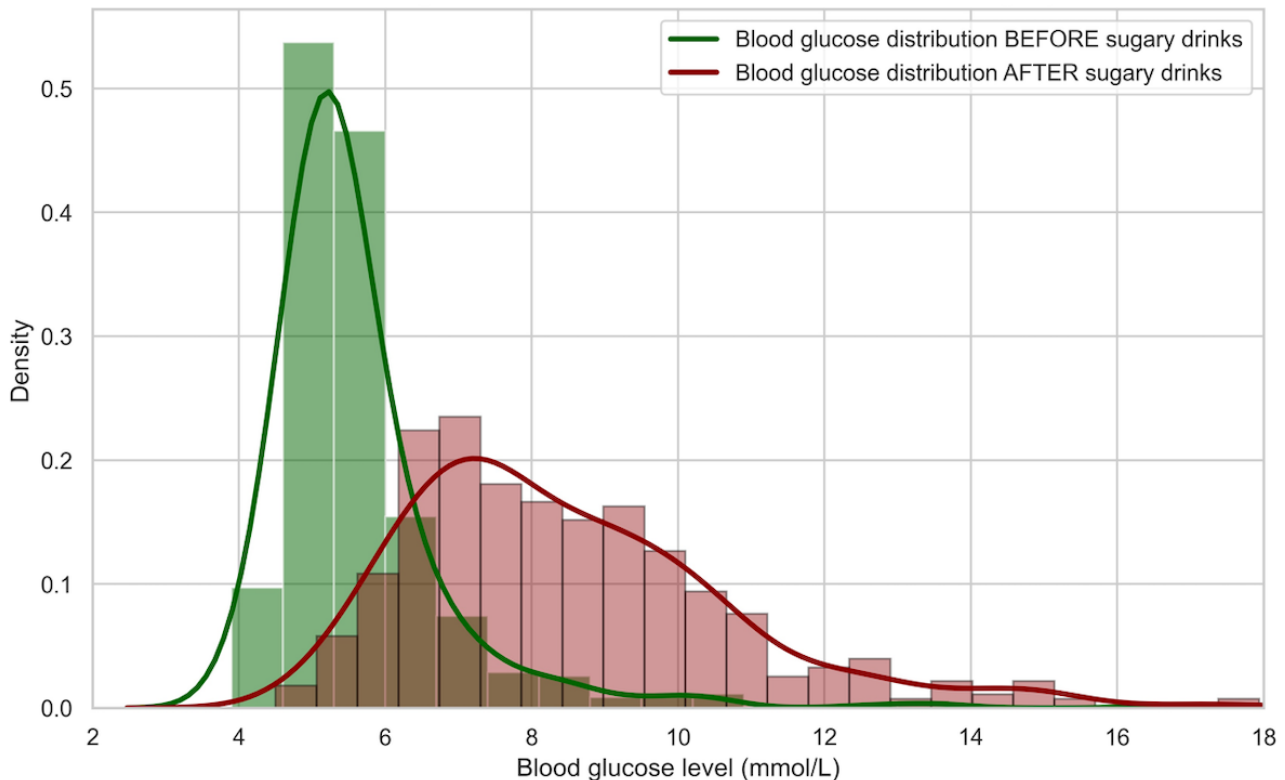
After consuming the sugary drink, 55.1% (266/483) of the participants had high blood glucose (≥ 7.8 mmol/L). The distribution of blood glucose levels before and after consuming the sugary drink is shown in Figure 2. A statistically significant difference was observed between the 2 distributions ($P < .001$).

Table 1. Description of participants (N=500).

Characteristics	Values
Demographic data	
Age (years), mean (SD); range	38.73 (10.61); 21-81
BMI (kg/m ²), mean (SD); range	24.4 (5.1); 16.3-71.1
Gender, n (%)	
Men	51 (10.2)
Women	449 (89.8)
Diabetes profile	
Family history of diabetes, n (%)	
Yes	157 (31.4)
No	343 (68.6)
Prediabetes, n (%)	
Yes	17 (3.4)
No	483 (96.6)
Diabetes, n (%)	
Yes	8 (1.6)
No	492 (98.4)
Gestational diabetes, n (%)	
Yes	21 (4.2)
No	428 (85.6)
N/A ^a	51 (10.2)

^aN/A: not applicable.

Figure 2. The distribution of ground-truth blood glucose levels before and after sugary drinks ($P < .001$).



Study Device

The Actxa Spark+ Series 2, a low-cost and commercially available wrist-worn wearable device, was used in this project. This multifunctional device, built for everyday activities, fitness, and preventive health monitoring, provided an adequate PPG signal quality at 50 Hz. The wearable device is equipped with advanced PPG technology that enables accurate and reliable measurement of heart rate (HR) and other physiological parameters. This is similar to the devices used in Singapore's nationwide health care campaigns, such as the National Steps Challenge. It is also worth noting that our proposed solution is device agnostic and can be easily integrated into other wearables with PPG capabilities, allowing for a scalable and cost-effective assessment of risk-based populations, including high-risk participants, participants with undiagnosed diabetes, and patients in need of primary prevention interventions.

Before Processing

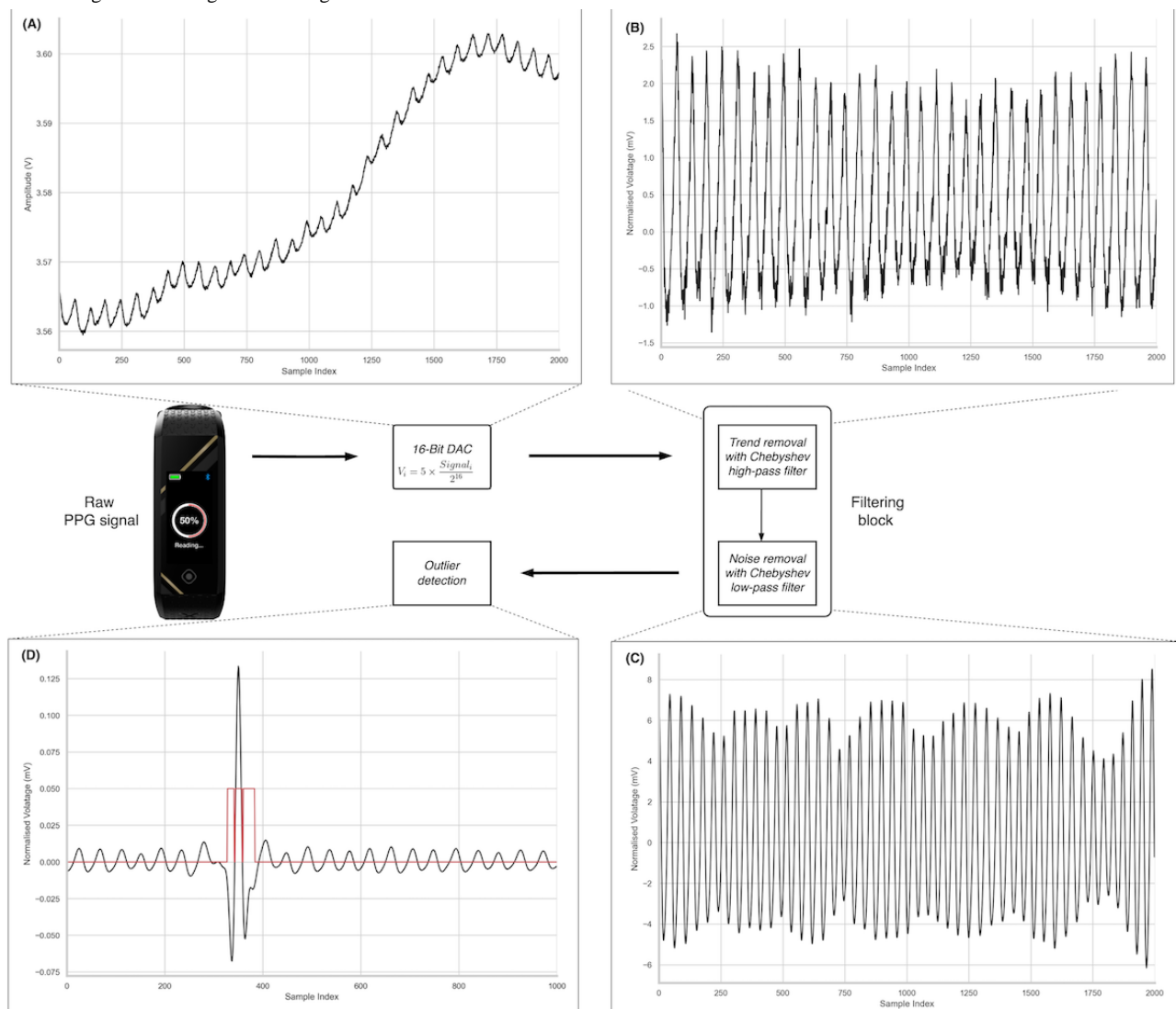
The raw PPG signal was collected using both wrist-worn wearables in 16-bit binary format. We first performed a digital-to-analog conversion using the following formula:

$$x = \frac{y - \min(y)}{\max(y) - \min(y)} \times (x_{\max} - x_{\min}) + x_{\min}$$

Liang et al [41] suggested that a fourth-order Chebyshev II filter provides an optimal processing performance for short PPG signals. Hence, we adopted the recommended filter design to remove low-frequency drift and high-frequency noise using a band-pass Chebyshev II filter. The proposed band-pass filter has a lower cut-off frequency of 0.3 Hz and an upper cut-off frequency of 4 Hz.

The filtered PPG signals still contain various forms of outliers, such as peaks with abnormally high amplitudes or distortions in the oscillating waveform, which can be caused by movement from the upper extremity or improper contact between the sensor and skin. Features derived from signals that possess outliers may not be accurate, so a z scores outlier detection with a cut-off value of 3 SDs of the mean was applied. The identified outliers or regions of outliers were replaced with a reasonable estimate via a nearest neighbor interpolation for the HRV feature extraction. Because PPG signals do not change drastically in such a short duration, this method is determined to be an appropriate approach to the problem. Furthermore, the number of outliers was minimal in our data set, and hence should not have affected the features that we generated later. The data preprocessing steps are illustrated in Figure 3.

Figure 3. Data preprocessing workflow. (A) Raw photoplethysmography (PPG) signal, (B) removal of the signal's moving trend using a Chebyshev high-pass filter, (C) use of a Chebyshev low-pass filter to eliminate high-frequency noise, and (D) final step involves outlier identification from the filtered PPG signal. DAC: digital-to-analog conversion.



Feature Extraction

Overview

The preprocessed data were suitable for generating reliable features, and a total of 248 features were generated. These features can be classified into seven categories: (1) HRV features, which encompass time domain, frequency domain, and nonlinear HRV features; (2) waveform features; (3) HR features; (4) energy measure features; (5) complexity measure features; (6) continuous wavelet transform (CWT) features; and (7) patient demographics. The complete set of features analyzed in this study is summarized in [Multimedia Appendix 1](#). However, these 248 feature candidates are not all relevant to the change in glucose level, and redundant features might cause prediction performance deterioration. The details of the feature-engineering and feature-selection process are discussed in the “Feature Selection” section.

HRV Features

HRV is the variation in time intervals between consecutive heartbeats and is widely used as a noninvasive physiological

biomarker of the autonomic nervous system response [42-44]. HRV provides a proxy to measure sympathetic nervous system (SNS) and parasympathetic nervous system (PNS) activity, which reflects the ability to respond to and recover from abrupt physical, psychological, and environmental changes [44-46]. As HR estimated at any given time represents the net effect of the neural output of the PNS, which slows HR, and SNS, which accelerates HR, HRV also detects imbalance in the autonomic nervous system resulting from over- or understimulation of SNS and PNS. Therefore, the fluctuation in HRV values provide useful insights into many clinical applications, such as mental stress, exercise and rehabilitation, cardiovascular fitness, pathological state, progression of chronic disease, and even predicting the onset of diseases [47-51]. Depending on the application, HRV features are usually extracted from an ultra-short-term (<5 min), short-term (approximately 5 min), or whole-day 24-hour time frame [52]. Most HRV features can be grouped under time-domain, frequency-domain, or nonlinear categories. In this project, most of the widely used HRV features were included in our analysis and were extracted using a

5-minute time frame. These HRV features are briefly explained in [Multimedia Appendix 1](#) using the feature indices (F1-F71).

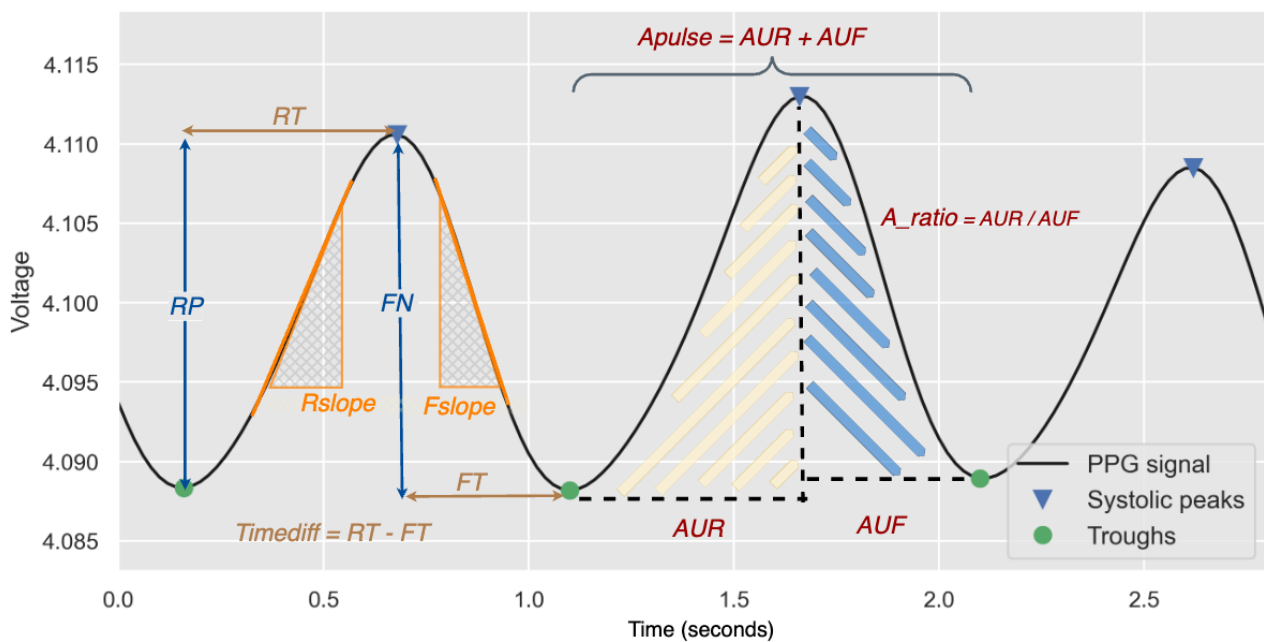
HR Features

Prior studies have noted the influence of impaired blood glucose on HR, especially resting HR [53,54]. Hence, HR was extracted by finding the number of peaks for every 10 seconds of the filtered PPG signal. The statistical features of the HR were then calculated and used as part of the feature inputs (F72-F81).

Wavelet Analysis

A considerable number of studies have applied wavelet transformation to analyze HRV data associated with a wide variety of health care applications. Earlier research has used features derived from CWT to predict blood glucose levels [55]. In this project, we applied CWT to the PPG signal using the Mexican Hat mother wavelet. The mean, SD, and maximum value of the resulting CWT matrix were included in the feature vector (F82-F84).

Figure 4. Definition of the photoplethysmography (PPG) waveform features. AUF: area under the falling edge; Apulse: area under a PPG wave; AUR: area under the rising edge; FN: magnitude of falling edge; Fslope: slope of falling edge; FT: fall time; RP: magnitude of rising edge; Rslope: slope of rising edge; RT: rise time.



Energy Measures

Several studies have used the energy features extracted from PPG signals to estimate blood glucose [34,59,60]. The Kaiser-Teager energy (KTE) operator and logarithmic energy are 2 commonly used methods to analyze the energy profile. These features were computed from a 5-second sliding window, as it ensures that the PPG signals within each window would be long enough to contain several heartbeats but short enough such that the wave amplitude changes are negligible.

The KTE operator is a well-known method for providing a time-frequency analysis of the instantaneous energy of the PPG signal from the amplitude and frequency. Using the implementation strategy explained by Monte-Moreno [34], we computed the energy profile of the PPG signal at each sliding

Waveform Features

Previous studies have reported that the characteristics of the PPG waveform extracted from healthy participants and participants with diabetes exhibited statistical differences [37,56]. Nirala et al [56] also suggested that the first and second eigenvalues derived from the first derivative of the PPG signal are the top features for identifying T2DM. In addition, several studies have revealed a functional relationship between the PPG signal and blood glucose levels [34,57]. Similarly, respiratory information can also be extracted from the PPG waveform [33,58]. However, PPG waveforms derived from signals using a wrist-worn PPG sensor often have a nondetectable diastolic peak and a dicrotic notch, unlike the signals collected using fingertip PPG.

Waveform features (F85-F196) derived from the PPG waveform were included in the feature set, and the definition of the waveform features is illustrated in [Figure 4](#).

window frame, and the KTE operator for the n -th frame was computed using the following equation:

$$KTE_n(i) = x_{frame}(i)^2 - x_{frame}(i+1) * x_{frame}(i-1), \text{ which holds for } i = 2, 3, \dots, (L_{frame} - 1) \quad (2)$$

Where x_{frame} is the filtered PPG signal within each sliding window frame.

The statistical metrics were computed for each frame, and the average of the metrics for the n th frame was then calculated and represented as F197 to F206.

To estimate the respiration rate from the PPG signal, we used the logarithmic energy value calculated at the frame level using the following equation:



Where x_{frame} is the filtered PPG signal within each sliding window frame.

The autoregressive model coefficients of order 7 were estimated using the Yule-Walker method, and the Python function *aryule* was used for this purpose. In addition, other statistical parameters were also computed (F207-F223).

Complexity Measures

Sample entropy (SampEn, F224) measures the unpredictability of physiological signals and is commonly used in HRV analysis [61]. The lower the SampEn, the more regular the signal.

SampEn can be defined after calculating the template vector ϕ^m that is the probability that 2 sequences will match for m points without allowing self-counting [62]:



Where m denotes the embedding dimension, tolerance r equals $0.1 * SD$, N denotes the number of data points, and C_i^m counts, within the tolerance resolution r , the number of matching blocks across different embedding dimensions.

SampEn is a tool used to analyze physiological time-series data, but it does not evaluate the complexity of the data at different time scales. Hence, we applied multiscale entropy (MSE) analysis on raw PPG signals to evaluate the hypothetical difference in signal complexity across various time scales for normoglycemia and elevated glucose levels. However, the scale factor was inversely proportionate to the number of data points. From our empirical results, we found that a minimum of 240 pulse waves were required to correctly compute the MSE values over all the timescale factors ($\tau=20$). We found that the sample entropy calculated from PPG signals during periods of elevated blood glucose was significantly higher than that of blood glucose in the normal range at timescale factors between 8 and 14 (τ). This information was then used to create features for the detection of elevated blood glucose levels. Each timescale factor between 8 and 14 was used as a separate feature. In addition, the mean of the adjacent timescale factors was derived to create additional features. These MSE features are represented in the feature vector with feature indices F225 to F244.

Results

Software

All experiments and analyses were performed using Python (version 3.9) and relevant libraries (Table 2). The final model was deployed on Amazon Web Services.

Table 2. A list of the software, and relevant libraries, along with the versions used.

Library	Version
Python	3.9.10
Imbalanced-learn	0.10.1
Joblib	1.2.0
Jupyter	1.0.0
Lightgbm	3.3.4
Matplotlib	3.6.2
Neurokit2	0.2.2
Nolds	0.5.2
Numba	0.53.1
Numpy	1.23.5
Pandas	1.5.2
Pillow	9.3.0
Scikit-learn	1.1.3
Scipy	1.8.0
Seaborn	0.12.1
Spectrum	0.8.1
Statsmodels	0.13.5
Xgboost	1.7.2

Feature Selection

Considering AI ethics and the practicality of implementing the algorithm, some demographic data, such as skin color, race, and personal lifestyle habits, were not used as inputs to the

models. However, other general personal characteristics associated with the risk of developing T2DM, such as age, gender, BMI, and family health history of diabetes, were added to the feature vector before the feature-selection process.

The redundant or irrelevant features might hinder the performance of the prediction model. To reduce the dimensionality of the input features, we applied an ensemble strategy that uses multiple feature-selection algorithms. This creates an optimal feature subset that minimizes the prediction error rate and is the most relevant for predicting the target variable. The ensemble feature-selection steps are summarized as follows:

- Six feature-selection methods, including ANOVA correlation coefficient, mutual information, dispersion ratio, recursive feature elimination, lasso regression, and Extreme Gradient Boosting, were used to choose the 30 best features independently.

- We combined the features obtained from each feature-selection method and ranked them using a majority vote approach to find the common features selected by more than 1 model.
- The highly correlated features were dropped from the selected feature subset.

In total, 12 features were selected from the entire feature set and ranked based on the results of the feature-selection strategy (Table 3). In our study, these selected features were the most sensitive predictors for capturing the characteristics of a participant's elevated blood glucose levels.

Table 3. The selected top features after the ensemble feature-selection method.

Rank	Feature
1	Welch_hf_rel
2	AR_hf_rel
3	A_FE_mean
4	A_ratio_mean
5	Age
6	A_Pulse_iqr
7	KTE_skew
8	LOG_std
9	BMI
10	MSE_sum_13_14
11	Family history
12	A_ratio_max
13	Gender ^a

^aNote that gender was not selected as a top feature in our feature-selection algorithm. However, it was previously identified as a sensitive predictor for T2DM, in which the prevalence of T2DM in men was higher than that in women [63]. This discrepancy could be attributed to the gender imbalance in the data set (men: 10.2%; women: 89.8%). Therefore, we included gender as one of the top features to provide a complete user profile for future investigation and development.

The selected features could be further divided into 4 main categories. Under the time-domain features, the selected features were the area under the PPG curves. A_FE_mean refers to the average area under the falling edge of each pulse (Figure 4). A_ratio refers to the ratio of the area under the rising edge to the area under the falling edge of each pulse (Figure 4), and both the average and maximum values were deemed relevant to the model's predictions. A_pulse_iqr refers to the IQR of the total area under each pulse (Figure 4). In the frequency domain, the selected features were the relative powers of the high-frequency bands in both the Welch power spectral density (PSD; Multimedia Appendix 1, F32-F44) and autoregressive PSD (Multimedia Appendix 1, F45-F57).

In the nonlinear domain, the selected features were either related to the energy or the complexity of the signal. LOG_std refers to the SD of log-energy entropy (equation 3), whereas KTE_skew refers to the skewness of the KTE energy measure for each sliding window (equation 2). Furthermore, the complexity feature that was selected was the sum of the MSE over 2 scales, 13 and 14.

Finally, the remaining selected features were demographic features that described the age and BMI of the participants, as well as if they had any family history of diabetes.

Machine Learning Model Performance

Seven widely used machine learning (ML) algorithms, including the naive Bayes classifier, K-nearest neighbors algorithm, logistic regression, random forest, SVM, XGB, and light gradient boosting machine, were trained with the selected features as inputs. We fine-tuned the hyperparameters of each model and validated their performance using the stratified 10-fold cross-validation method. We adopted multiple regularization techniques across various models to prevent overfitting during the model training. Six evaluation metrics, accuracy, sensitivity, specificity, precision, geometric mean (G-mean), and *F* score, were used to evaluate the model's performance, as accuracy alone cannot provide a comprehensive examination of model performance due to data imbalance. The G-mean and *F* score are critical evaluation criteria to assess the

models' performance, as they are robust to significant label imbalance.

The prediction results from each model are reported as the mean and SD of the evaluation metrics, and Table 4 shows the

summary of the results. SVM with the radial basis function kernel showed the best prediction performance with an average accuracy of 84.7%, a sensitivity of 81.05%, a specificity of 88.35%, and a precision of 87.51%. In particular, the average G-mean was 84.54% and *F* score was 84.03%.

Table 4. The prediction results obtained from 10-fold cross-validation using various machine learning models.

Model	Accuracy		Sensitivity		Specificity		Precision		Geometric mean		<i>F</i> score	
	μ	σ	μ	σ	μ	σ	μ	σ	μ	σ	μ	σ
NB ^a	60.51	4.63	66.17	7.44	54.87	5.78	59.43	4.12	60.08	4.6	62.51	5.19
KNN ^b	76.7	3	90.45	4.30	62.94	4.15	70.97	2.47	75.4	3.09	79.5	2.68
LR ^c	63.1	4.65	64.56	7.07	61.66	4.30	62.65	4.16	63	4.67	63.52	5.37
RF ^d	76.76	5.73	76.84	8.18	76.69	6.42	76.81	6.08	76.64	5.72	76.68	6.23
SVM ^e	84.7	4.14	81.05	6.77	88.34	4.19	87.51	4.26	84.54	4.18	84.03	4.58
XGB ^f	78.06	4.91	77	6.58	79.12	4.98	78.7	4.88	78	4.89	77.77	5.15
LGBM ^g	77.9	3.98	75.54	7.36	80.27	4.45	79.35	4.1	77.74	4.07	77.24	4.81

^aNB: naive Bayes.

^bKNN: K-nearest neighbors.

^cLR: logistic regression.

^dRF: random forest.

^eSVM: support vector machine.

^fXGB: Extreme Gradient Boosting.

^gLGBM: light gradient boosting machine.

Model Interpretation

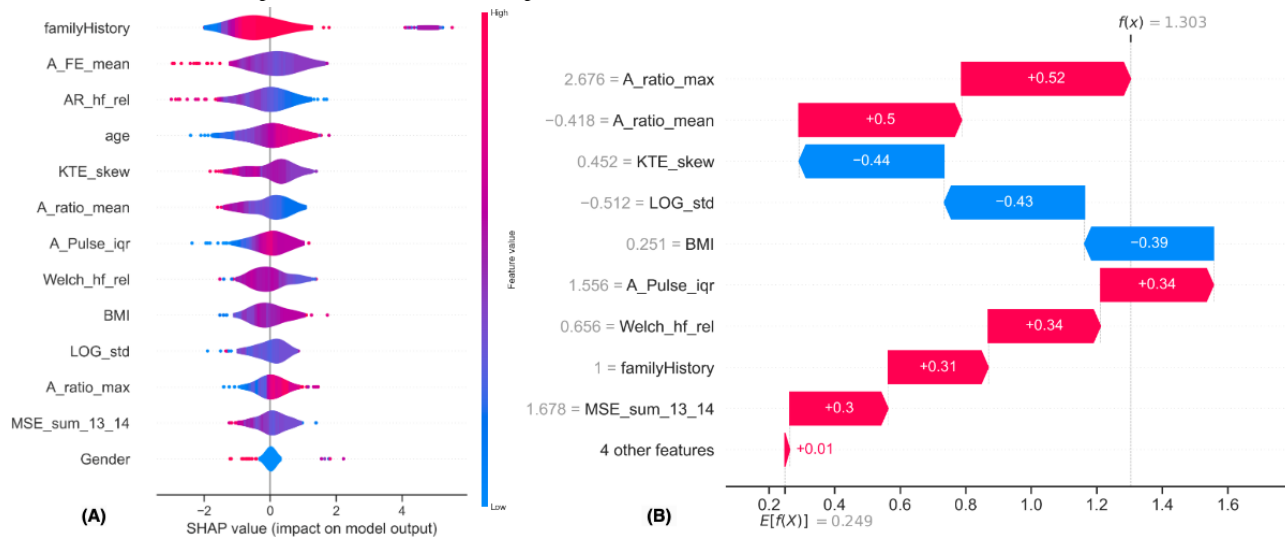
The use of deep learning in the medical and health care domain has shown great potential for solving a range of problems, such as detecting specific symptoms or abnormalities [64,65]. However, the interpretability of deep learning models remains a significant challenge, and it is often difficult for clinicians to trust the decisions made using a black-box system. The lack of model interpretability also raises ethical concerns, particularly when the decision fails. Furthermore, our current data set is considerably small (500 participants) compared with typical deep learning models in other domains, which are trained with thousands of data points. Deep learning models are known to perform well with a larger data set and fail to learn meaningful representations when there is a lack of data [66]. Therefore, we did not investigate the use of deep learning in this study.

As the proposed ML model is designed to complement the existing diabetes detection solution and is relatively new to the clinical community, the features selected in the previous section must be interpretable and exhibit a certain level of agreement with existing findings. A family history of diabetes, being male, being aged ≥ 45 years, and having an increased BMI have been

identified as major risk factors in the literature for developing prediabetes or T2DM [63,67,68]. These 4 risk factors were part of the selected predictors, and this paper provides a preliminary attempt to explain how the selected predictors contribute to detecting elevated blood glucose using the Shapley additive explanations (SHAP) framework. SHAP is a game theoretical approach that provides global and local explanations of the association between the ML output and input features [69].

Figure 5A illustrates the SHAP values of each feature across all the predictions from the training set. The features were ranked by their mean SHAP values, with larger values shown in red and smaller values shown in blue. The beeswarm plot revealed that a family history of diabetes, increasing age, and higher BMI are associated with a higher probability of elevated blood glucose levels. These observations are consistent with previous research and demonstrate that the ML algorithm has successfully captured the relationship between these features and elevated blood glucose levels. In addition, other proposed features showed varying levels of impact on the model's output. However, the gender feature did not have any apparent effect on the model's predictions.

Figure 5. The Shapley additive explanations (SHAP) plots indicate the association between the selected features and their impact on the predicted outcome. (A) SHAP beeswarm plot and (B) SHAP waterfall plot.



In Figure 5B, each row in the plot shows how the contributions of different features move the output of the model from the expected value ($E[f(x)]$) to the actual prediction output $f(x)$ for a single sample with a positive class prediction (blood glucose level ≥ 7.8 mmol/L) in the test set. The expected value, $E[f(x)]$, is determined using the entire training data set. As expected, most features provide positive SHAP values in this sample, which collectively push the model's output toward the correct prediction. However, this specific test participant's BMI was in the healthy range, which pushed the model's output toward the normal class and might have resulted in a false negative prediction. This indicates that relying on a single feature or demographic data alone may not provide an accurate prediction of blood glucose levels.

Using the SHAP values, we can understand the model's overall behaviors and how features affect the output positively or negatively, which can help improve the prediction model in the future.

Assessment of the Elevated Blood Glucose Levels From Multiple Measurements

Generally, diagnostic tests are not highly sensitive and highly specific. Therefore, repeated measurements of the wrist-worn wearable device were combined and assessed in an optimum fashion to maximize sensitivity, specificity, and precision.

Consecutive measures of blood glucose were combined in parallel using the "AND" and "OR" rules to assist in the detection of elevated blood glucose measurement levels. The "OR" rule increases the overall sensitivity, and the "AND" rule increases the overall specificity, which is greater than that of either test alone [70].

Discussion

Principal Findings

While the health care landscape is changing, the rapidly aging society and the need for improved population health outcomes call for new models of care to effectively prevent the onset and delay the progression of chronic diseases. Furthermore,

short-term health behaviors contribute significantly toward long-term health outcomes, while unattended and frequent glucose spikes might result in prediabetes and eventually diabetes. The availability of noninvasive and device-agnostic blood glucose detection solutions will allow for more frequent and better monitoring of blood glucose levels, thereby reducing the risk of developing T2DM. This study demonstrates that a noninvasive method of assessing diabetes risk using PPG is a viable option to provide a cheaper and accessible modality for the population-wide screening of blood glucose levels. This population-based screening would allow for the earlier detection of DM in the population, especially among those individuals who are unaware of their elevated blood glucose levels. Hence, timely and appropriate lifestyle advice and medical interventions can be provided to prevent diabetes complications. This will subsequently reduce the health care burden for both the individual and the society.

BGEM is a cloud-based solution that can frequently monitor multiple digital biomarkers with minimal disruption to daily life. Developed using the advanced ML operations practice, BGEM can be easily scaled to meet the increasing demand for health care services. The solution includes a user-friendly mobile app that can screen a large population to identify high-risk individuals, people with undiagnosed diabetes, and those who require primary prevention intervention. It also provides timely feedback to users through the app, informing them of their diabetes risk and providing targeted, actionable insights to empower them to take a proactive approach to monitor their glucose levels.

Limitations

Our pilot study has certain limitations. Since fasting blood glucose measurements were excluded and the criteria to define normal and abnormal levels under fasting conditions differed from our current cut-off, we must refrain from definitively concluding that our model is applicable to fasting conditions. Regarding gender, our feature-selection model did not specifically incorporate it, and our analysis using SHAP demonstrated that gender exerted minimal influence on model predictions. Moreover, all analyses were adjusted for the

covariate gender, as required. Therefore, we considered gender to have a limited impact and is not a primary limitation of our findings. To address these limitations, we are actively planning the subsequent phase of data collection. This phase will involve collecting fasting blood glucose measurements in a primary care setting, also allowing for a more balanced gender distribution. More importantly, we could expand our participant pool to encompass participants with prediabetes and diabetes. By addressing these gaps, we aimed to offer a more comprehensive and robust assessment of our model's applicability and effectiveness.

There was no longitudinal follow-up of the participants. External validation of our model on an independent sample must be undertaken to further assess the detection accuracy and generalizability of the results. Nevertheless, as a preliminary investigation, the potential implications of our findings are significant as they might offer a means to identify previously undiagnosed prediabetes or diabetes cases at the population level. We anticipate that our study will serve as a foundational stepping stone, paving the way for more comprehensive diabetes research using AI and wearable devices. To the best of our knowledge, there is no publicly available data set that systematically examines the relationship between PPG data and blood glucose levels. Acquiring a substantial volume of data is imperative, encompassing a diverse and representative sample

spanning the entire spectrum of glucose values and incorporating relevant sociodemographic factors. Such comprehensive data can be obtained through a collaborative effort involving research institutions and industry partners while ensuring strict adherence to local ethical considerations and data privacy regulations.

We demonstrated that the cloud-based ML model can detect elevated blood glucose levels, where consecutive measurements can be combined in an optimal manner to provide high sensitivity, specificity, and precision. However, further research is required to address these limitations.

Conclusions

In this study, we performed sophisticated feature engineering and found that the features derived from the MSE analysis of PPG signals effectively detect blood glucose changes. We will discuss this set of novel features in detail in a separate paper. To reduce bias and evaluate the generalizability of the model, we used a 10-fold cross-validation to assess its performance. The SVM with the radial basis function model performed the best, with an average accuracy of 84.7%, a G-mean of 84.54%, and an *F* score of 84.03%. Previous models were developed using smaller samples and have lower model performance measures. Our model was developed with a larger sample of 500 participants, and most participants were assessed before and after the consumption of a sugary drink. It also achieved better detection accuracy.

Acknowledgments

This research was sponsored by Actxa Pte Ltd, but data collection was performed independently at KK Women's and Children's Hospital, Singapore.

Authors' Contributions

BS contributed to the study design, conducted the data analysis and experiments, developed the algorithms and models, and drafted the manuscript. SSD designed the study, performed the statistical data analysis, and drafted the manuscript. JW was responsible for model deployment and developed the data pipeline infrastructure. CC, MTC, KCSL, and FFI contributed to data collection. NWCL, EZ, KYL, and VP assisted with the development of the algorithms and supported data collection. AWHL performed data analysis and edited the manuscript. MS contributed to the study design and supervised the study. JC, S-CY, and AT supervised the study. SBA contributed to study design and supervised the study. All authors have reviewed the manuscript.

Conflicts of Interest

The authors would like to disclose that BS, MS, JW, AWHL, and JC are employed by Actxa Pte Ltd. The authors have an approved plan for managing any potential conflicts arising from employment. SBA and SSD are on the advisory board of Actxa Pte Ltd. All other authors declare that they have no conflicts of interest.

Multimedia Appendix 1

Features summary.

[[DOCX File, 21 KB - ai_v2i1e48340_app1.docx](#)]

References

1. National Diabetes Data Group. Classification and diagnosis of diabetes mellitus and other categories of glucose intolerance. *Diabetes* 1979 Dec 1;28(12):1039-1057. [doi: [10.2337/diab.28.12.1039](https://doi.org/10.2337/diab.28.12.1039)] [Medline: [510803](https://pubmed.ncbi.nlm.nih.gov/510803/)]
2. Kerner W, Brückel J. Definition, classification and diagnosis of diabetes mellitus. *Exp Clin Endocrinol Diabetes* 2014 Jul 11;122(7):384-386. [doi: [10.1055/s-0034-1366278](https://doi.org/10.1055/s-0034-1366278)] [Medline: [25014088](https://pubmed.ncbi.nlm.nih.gov/25014088/)]
3. American Diabetes Association. 2. Classification and diagnosis of diabetes: standards of medical care in diabetes—2020. *Dia Care* 2019 Dec 20;43(Supplement 1):S14-S31 [[FREE Full text](#)] [doi: [10.2337/dc20-s002](https://doi.org/10.2337/dc20-s002)]

4. IDF diabetes atlas 10th edition. International Diabetes Federation. 2021. URL: https://diabetesatlas.org/idfawp/resource-files/2021/07/IDF_Atlas_10th_Edition_2021.pdf [accessed 2023-06-04]
5. Emerging Risk Factors Collaboration, Sarwar N, Gao P, Seshasai SR, Gobin R, Kaptoge S, et al. Diabetes mellitus, fasting blood glucose concentration, and risk of vascular disease: a collaborative meta-analysis of 102 prospective studies. *Lancet* 2010 Jun 26;375(9733):2215-2222 [FREE Full text] [doi: [10.1016/S0140-6736\(10\)60484-9](https://doi.org/10.1016/S0140-6736(10)60484-9)] [Medline: [20609967](https://pubmed.ncbi.nlm.nih.gov/20609967/)]
6. Lin X, Xu Y, Pan X, Xu J, Ding Y, Sun X, et al. Global, regional, and national burden and trend of diabetes in 195 countries and territories: an analysis from 1990 to 2025. *Sci Rep* 2020 Sep 08;10(1):14790 [FREE Full text] [doi: [10.1038/s41598-020-71908-9](https://doi.org/10.1038/s41598-020-71908-9)] [Medline: [32901098](https://pubmed.ncbi.nlm.nih.gov/32901098/)]
7. Li S, Wang J, Zhang B, Li X, Liu Y. Diabetes mellitus and cause-specific mortality: a population-based study. *Diabetes Metab J* 2019 Jun;43(3):319-341 [FREE Full text] [doi: [10.4093/dmj.2018.0060](https://doi.org/10.4093/dmj.2018.0060)] [Medline: [31210036](https://pubmed.ncbi.nlm.nih.gov/31210036/)]
8. Saran R, Li Y, Robinson B, Ayanian J, Balkrishnan R, Bragg-Gresham J, et al. US renal data system 2014 annual data report: epidemiology of kidney disease in the United States. *Am J Kidney Dis* 2015 Jul;66(1):S1-305 [FREE Full text] [doi: [10.1053/j.ajkd.2015.05.001](https://doi.org/10.1053/j.ajkd.2015.05.001)] [Medline: [26111994](https://pubmed.ncbi.nlm.nih.gov/26111994/)]
9. Lau LH, Lew J, Borschmann K, Thijs V, Ekinici EI. Prevalence of diabetes and its effects on stroke outcomes: a meta-analysis and literature review. *J Diabetes Investig* 2019 May;10(3):780-792 [FREE Full text] [doi: [10.1111/jdi.12932](https://doi.org/10.1111/jdi.12932)] [Medline: [30220102](https://pubmed.ncbi.nlm.nih.gov/30220102/)]
10. Sun H, Saeedi P, Karuranga S, Pinkepank M, Ogurtsova K, Duncan BB, et al. IDF diabetes atlas: global, regional and country-level diabetes prevalence estimates for 2021 and projections for 2045. *Diabetes Res Clin Pract* 2022 Jan;183:109119. [doi: [10.1016/j.diabres.2021.109119](https://doi.org/10.1016/j.diabres.2021.109119)] [Medline: [34879977](https://pubmed.ncbi.nlm.nih.gov/34879977/)]
11. Huang Y, Cai X, Mai W, Li M, Hu Y. Association between prediabetes and risk of cardiovascular disease and all cause mortality: systematic review and meta-analysis. *BMJ* 2016 Nov 23;355:i5953 [FREE Full text] [doi: [10.1136/bmj.i5953](https://doi.org/10.1136/bmj.i5953)] [Medline: [27881363](https://pubmed.ncbi.nlm.nih.gov/27881363/)]
12. Wong M, Gu K, Heng D, Chew SK, Chew LS, Tai ES. The Singapore impaired glucose tolerance follow-up study: does the ticking clock go backward as well as forward? *Diabetes Care* 2003 Nov;26(11):3024-3030 [FREE Full text] [doi: [10.2337/diacare.26.11.3024](https://doi.org/10.2337/diacare.26.11.3024)] [Medline: [14578234](https://pubmed.ncbi.nlm.nih.gov/14578234/)]
13. Tabák AG, Herder C, Rathmann W, Brunner EJ, Kivimäki M. Prediabetes: a high-risk state for diabetes development. *The Lancet* 2012 Jun;379(9833):2279-2290 [FREE Full text] [doi: [10.1016/s0140-6736\(12\)60283-9](https://doi.org/10.1016/s0140-6736(12)60283-9)]
14. US Preventive Services Task Force, Davidson K, Barry MJ, Mangione CM, Cabana M, Caughey AB, et al. Screening for prediabetes and type 2 diabetes: US preventive services task force recommendation statement. *JAMA* 2021 Aug 24;326(8):736-743 [FREE Full text] [doi: [10.1001/jama.2021.12531](https://doi.org/10.1001/jama.2021.12531)] [Medline: [34427594](https://pubmed.ncbi.nlm.nih.gov/34427594/)]
15. Manne-Goehler J, Geldsetzer P, Agoudavi K, Andall-Brereton G, Aryal KK, Bicaba BW, et al. Health system performance for people with diabetes in 28 low- and middle-income countries: a cross-sectional study of nationally representative surveys. *PLoS Med* 2019 Mar 1;16(3):e1002751 [FREE Full text] [doi: [10.1371/journal.pmed.1002751](https://doi.org/10.1371/journal.pmed.1002751)] [Medline: [30822339](https://pubmed.ncbi.nlm.nih.gov/30822339/)]
16. Misra A, Gopalan H, Jayawardena R, Hills AP, Soares M, Reza-Albarrán AA, et al. Diabetes in developing countries. *J Diabetes* 2019 Mar 12;11(7):522-539. [doi: [10.1111/1753-0407.12913](https://doi.org/10.1111/1753-0407.12913)] [Medline: [30864190](https://pubmed.ncbi.nlm.nih.gov/30864190/)]
17. García-Molina L, Lewis-Mikhael A, Riquelme-Gallego B, Cano-Ibáñez N, Oliveras-López MJ, Bueno-Cavanillas A. Improving type 2 diabetes mellitus glycaemic control through lifestyle modification implementing diet intervention: a systematic review and meta-analysis. *Eur J Nutr* 2020 Jun;59(4):1313-1328 [FREE Full text] [doi: [10.1007/s00394-019-02147-6](https://doi.org/10.1007/s00394-019-02147-6)] [Medline: [31781857](https://pubmed.ncbi.nlm.nih.gov/31781857/)]
18. O'Donoghue G, O'Sullivan C, Corridan I, Daly J, Finn R, Melvin K, et al. Lifestyle interventions to improve glycaemic control in adults with type 2 diabetes living in low-and-middle income countries: a systematic review and meta-analysis of Randomized Controlled Trials (RCTs). *Int J Environ Res Public Health* 2021 Jun 10;18(12):6273 [FREE Full text] [doi: [10.3390/ijerph18126273](https://doi.org/10.3390/ijerph18126273)] [Medline: [34200592](https://pubmed.ncbi.nlm.nih.gov/34200592/)]
19. Tusso P. Prediabetes and lifestyle modification: time to prevent a preventable disease. *Perm J* 2014;18(3):88-93 [FREE Full text] [doi: [10.7812/TPP/14-002](https://doi.org/10.7812/TPP/14-002)] [Medline: [25102521](https://pubmed.ncbi.nlm.nih.gov/25102521/)]
20. Bansal N. Prediabetes diagnosis and treatment: a review. *World J Diabetes* 2015 Mar 15;6(2):296-303 [FREE Full text] [doi: [10.4239/wjd.v6.i2.296](https://doi.org/10.4239/wjd.v6.i2.296)] [Medline: [25789110](https://pubmed.ncbi.nlm.nih.gov/25789110/)]
21. Magkos F, Hjorth MF, Astrup A. Diet and exercise in the prevention and treatment of type 2 diabetes mellitus. *Nat Rev Endocrinol* 2020 Oct 20;16(10):545-555 [FREE Full text] [doi: [10.1038/s41574-020-0381-5](https://doi.org/10.1038/s41574-020-0381-5)] [Medline: [32690918](https://pubmed.ncbi.nlm.nih.gov/32690918/)]
22. Simmons RK, Griffin SJ, Lauritzen T, Sandbæk A. Effect of screening for type 2 diabetes on risk of cardiovascular disease and mortality: a controlled trial among 139,075 individuals diagnosed with diabetes in Denmark between 2001 and 2009. *Diabetologia* 2017 Nov;60(11):2192-2199 [FREE Full text] [doi: [10.1007/s00125-017-4299-y](https://doi.org/10.1007/s00125-017-4299-y)] [Medline: [28831539](https://pubmed.ncbi.nlm.nih.gov/28831539/)]
23. Heinemann L. Finger pricking and pain: a never ending story. *J Diabetes Sci Technol* 2008 Sep 01;2(5):919-921 [FREE Full text] [doi: [10.1177/193229680800200526](https://doi.org/10.1177/193229680800200526)] [Medline: [19885279](https://pubmed.ncbi.nlm.nih.gov/19885279/)]
24. Hambling CE, Seidu SI, Davies MJ, Khunti K. Older people with Type 2 diabetes, including those with chronic kidney disease or dementia, are commonly overtreated with sulfonylurea or insulin therapies. *Diabet Med* 2017 Sep;34(9):1219-1227 [FREE Full text] [doi: [10.1111/dme.13380](https://doi.org/10.1111/dme.13380)] [Medline: [28498634](https://pubmed.ncbi.nlm.nih.gov/28498634/)]

25. Mattishent K, Lane K, Salter C, Dhatariya K, May HM, Neupane S, et al. Continuous glucose monitoring in older people with diabetes and memory problems: a mixed-methods feasibility study in the UK. *BMJ Open* 2019 Nov 18;9(11):e032037 [FREE Full text] [doi: [10.1136/bmjopen-2019-032037](https://doi.org/10.1136/bmjopen-2019-032037)] [Medline: [31740472](https://pubmed.ncbi.nlm.nih.gov/31740472/)]
26. Vettoretti M, Cappon G, Acciaroli G, Facchinetti A, Sparacino G. Continuous glucose monitoring: current use in diabetes management and possible future applications. *J Diabetes Sci Technol* 2018 Sep 22;12(5):1064-1071 [FREE Full text] [doi: [10.1177/1932296818774078](https://doi.org/10.1177/1932296818774078)] [Medline: [29783897](https://pubmed.ncbi.nlm.nih.gov/29783897/)]
27. Funtanilla VD, Candidate P, Caliendo T, Hilas O. Continuous glucose monitoring: a review of available systems. *P T* 2019 Sep;44(9):550-553 [FREE Full text] [Medline: [31485150](https://pubmed.ncbi.nlm.nih.gov/31485150/)]
28. Robertson SL, Shaughnessy AF, Slawson DC. Continuous glucose monitoring in type 2 diabetes is not ready for widespread adoption. *Am Fam Physician* 2020 Jun 01;101(11):646 [FREE Full text] [Medline: [32463633](https://pubmed.ncbi.nlm.nih.gov/32463633/)]
29. Sabry F, Eltaras T, Labda W, Alzoubi K, Malluhi Q. Machine learning for healthcare wearable devices: the big picture. *J Healthc Eng* 2022 Apr 18;2022:4653923 [FREE Full text] [doi: [10.1155/2022/4653923](https://doi.org/10.1155/2022/4653923)] [Medline: [35480146](https://pubmed.ncbi.nlm.nih.gov/35480146/)]
30. Patel S, Park H, Bonato P, Chan L, Rodgers M. A review of wearable sensors and systems with application in rehabilitation. *J Neuroeng Rehabil* 2012 Apr 20;9(1):21 [FREE Full text] [doi: [10.1186/1743-0003-9-21](https://doi.org/10.1186/1743-0003-9-21)] [Medline: [22520559](https://pubmed.ncbi.nlm.nih.gov/22520559/)]
31. Rodgers MM, Alon G, Pai VM, Conroy RS. Wearable technologies for active living and rehabilitation: current research challenges and future opportunities. *J Rehabil Assist Technol Eng* 2019 Apr 26;6:2055668319839607 [FREE Full text] [doi: [10.1177/2055668319839607](https://doi.org/10.1177/2055668319839607)] [Medline: [31245033](https://pubmed.ncbi.nlm.nih.gov/31245033/)]
32. Xie Y, Lu L, Gao F, He SJ, Zhao HJ, Fang Y, et al. Integration of artificial intelligence, blockchain, and wearable technology for chronic disease management: a new paradigm in smart healthcare. *Curr Med Sci* 2021 Dec;41(6):1123-1133 [FREE Full text] [doi: [10.1007/s11596-021-2485-0](https://doi.org/10.1007/s11596-021-2485-0)] [Medline: [34950987](https://pubmed.ncbi.nlm.nih.gov/34950987/)]
33. Iqbal SM, Mahgoub I, Du E, Leavitt MA, Asghar W. Advances in healthcare wearable devices. *Npj Flex Electron* 2021 Apr 12;5(1):9 [FREE Full text] [doi: [10.1038/s41528-021-00107-x](https://doi.org/10.1038/s41528-021-00107-x)]
34. Monte-Moreno E. Non-invasive estimate of blood glucose and blood pressure from a photoplethysmograph by means of machine learning techniques. *Artif Intell Med* 2011 Oct;53(2):127-138 [FREE Full text] [doi: [10.1016/j.artmed.2011.05.001](https://doi.org/10.1016/j.artmed.2011.05.001)] [Medline: [21696930](https://pubmed.ncbi.nlm.nih.gov/21696930/)]
35. Rodin D, Kirby M, Sedogin N, Shapiro Y, Pinhasov A, Kreinin A. Comparative accuracy of optical sensor-based wearable system for non-invasive measurement of blood glucose concentration. *Clin Biochem* 2019 Mar;65:15-20 [FREE Full text] [doi: [10.1016/j.clinbiochem.2018.12.014](https://doi.org/10.1016/j.clinbiochem.2018.12.014)] [Medline: [30629956](https://pubmed.ncbi.nlm.nih.gov/30629956/)]
36. Zilberstein G, Zilberstein R, Maor U, Righetti PG. Noninvasive wearable sensor for indirect glucometry. *Electrophoresis* 2018 Sep;39(18):2344-2350 [FREE Full text] [doi: [10.1002/elps.201700424](https://doi.org/10.1002/elps.201700424)] [Medline: [29607521](https://pubmed.ncbi.nlm.nih.gov/29607521/)]
37. Zhang G, Mei Z, Zhang Y, Ma X, Lo B, Chen D, et al. A noninvasive blood glucose monitoring system based on smartphone PPG signal processing and machine learning. *IEEE Trans Industr Inform* 2020 Nov;16(11):7209-7218 [FREE Full text] [doi: [10.1109/tii.2020.2975222](https://doi.org/10.1109/tii.2020.2975222)]
38. Challoner AV, Ramsay CA. A photoelectric plethysmograph for the measurement of cutaneous blood flow. *Phys Med Biol* 1974 May;19(3):317-328 [FREE Full text] [doi: [10.1088/0031-9155/19/3/003](https://doi.org/10.1088/0031-9155/19/3/003)] [Medline: [4445210](https://pubmed.ncbi.nlm.nih.gov/4445210/)]
39. Zhao D, Sun Y, Wan S, Wang F. SFST: a robust framework for heart rate monitoring from photoplethysmography signals during physical activities. *Biomed Signal Process Control* 2017 Mar;33:316-324 [FREE Full text] [doi: [10.1016/j.bspc.2016.12.005](https://doi.org/10.1016/j.bspc.2016.12.005)]
40. Schroeder EB, Chambless LE, Liao D, Prineas RJ, Evans GW, Rosamond WD, et al. Diabetes, glucose, insulin, and heart rate variability: the Atherosclerosis Risk in Communities (ARIC) study. *Diabetes Care* 2005 Mar;28(3):668-674. [doi: [10.2337/diacare.28.3.668](https://doi.org/10.2337/diacare.28.3.668)] [Medline: [15735206](https://pubmed.ncbi.nlm.nih.gov/15735206/)]
41. Liang Y, Elgendi M, Chen Z, Ward R. An optimal filter for short photoplethysmogram signals. *Sci Data* 2018 May 01;5(1):180076 [FREE Full text] [doi: [10.1038/sdata.2018.76](https://doi.org/10.1038/sdata.2018.76)] [Medline: [29714722](https://pubmed.ncbi.nlm.nih.gov/29714722/)]
42. van Ravenswaaij-Arts CM, Kollée LA, Hopman JC, Stoelinga GB, van Geijn HP. Heart rate variability. *Ann Intern Med* 1993 Mar 15;118(6):436-447 [FREE Full text] [doi: [10.7326/0003-4819-118-6-199303150-00008](https://doi.org/10.7326/0003-4819-118-6-199303150-00008)] [Medline: [8439119](https://pubmed.ncbi.nlm.nih.gov/8439119/)]
43. Xhyheri B, Manfrini O, Mazzolini M, Pizzi C, Bugiardini R. Heart rate variability today. *Prog Cardiovasc Dis* 2012 Nov;55(3):321-331 [FREE Full text] [doi: [10.1016/j.pcad.2012.09.001](https://doi.org/10.1016/j.pcad.2012.09.001)] [Medline: [23217437](https://pubmed.ncbi.nlm.nih.gov/23217437/)]
44. Thomas BL, Claassen N, Becker P, Viljoen M. Validity of commonly used heart rate variability markers of autonomic nervous system function. *Neuropsychobiology* 2019 Feb 5;78(1):14-26. [doi: [10.1159/000495519](https://doi.org/10.1159/000495519)] [Medline: [30721903](https://pubmed.ncbi.nlm.nih.gov/30721903/)]
45. Obrist PA. *Cardiovascular Psychophysiology: A Perspective*. New York, NY: Springer; 1981.
46. Singh N, Moneghetti KJ, Christle JW, Hadley D, Plews D, Froelicher V. Heart rate variability: an old metric with new meaning in the era of using mHealth technologies for health and exercise training guidance. Part one: physiology and methods. *Arrhythm Electrophysiol Rev* 2018 Aug;7(3):193-198 [FREE Full text] [doi: [10.15420/aer.2018.27.2](https://doi.org/10.15420/aer.2018.27.2)] [Medline: [30416733](https://pubmed.ncbi.nlm.nih.gov/30416733/)]
47. Prinsloo GE, Rauch HL, Derman WE. A brief review and clinical application of heart rate variability biofeedback in sports, exercise, and rehabilitation medicine. *Phys Sportsmed* 2014 May 13;42(2):88-99 [FREE Full text] [doi: [10.3810/psm.2014.05.2061](https://doi.org/10.3810/psm.2014.05.2061)] [Medline: [24875976](https://pubmed.ncbi.nlm.nih.gov/24875976/)]
48. Billman GE, Huikuri HV, Sacha J, Trimmel K. An introduction to heart rate variability: methodological considerations and clinical applications. *Front Physiol* 2015 Feb 25;6:55 [FREE Full text] [doi: [10.3389/fphys.2015.00055](https://doi.org/10.3389/fphys.2015.00055)] [Medline: [25762937](https://pubmed.ncbi.nlm.nih.gov/25762937/)]

49. Kim HG, Cheon EJ, Bai DS, Lee YH, Koo BH. Stress and heart rate variability: a meta-analysis and review of the literature. *Psychiatry Investig* 2018 Mar;15(3):235-245 [FREE Full text] [doi: [10.30773/pi.2017.08.17](https://doi.org/10.30773/pi.2017.08.17)] [Medline: [29486547](https://pubmed.ncbi.nlm.nih.gov/29486547/)]
50. Taye GT, Hwang HJ, Lim KM. Application of a convolutional neural network for predicting the occurrence of ventricular tachyarrhythmia using heart rate variability features. *Sci Rep* 2020 Apr 21;10(1):6769 [FREE Full text] [doi: [10.1038/s41598-020-63566-8](https://doi.org/10.1038/s41598-020-63566-8)] [Medline: [32317680](https://pubmed.ncbi.nlm.nih.gov/32317680/)]
51. Mosley E, Laborde S. A scoping review of heart rate variability in sport and exercise psychology. *Int Rev Sport Exerc Psychol* 2022 Jul 07;1-75 [FREE Full text] [doi: [10.1080/1750984x.2022.2092884](https://doi.org/10.1080/1750984x.2022.2092884)]
52. Shaffer F, Ginsberg JP. An overview of heart rate variability metrics and norms. *Front Public Health* 2017 Sep 28;5:258 [FREE Full text] [doi: [10.3389/fpubh.2017.00258](https://doi.org/10.3389/fpubh.2017.00258)] [Medline: [29034226](https://pubmed.ncbi.nlm.nih.gov/29034226/)]
53. Valensi P, Extramiana F, Lange C, Cailleau M, Haggui A, Maison Blanche P, et al. Influence of blood glucose on heart rate and cardiac autonomic function. The DESIR study. *Diabet Med* 2011 Apr;28(4):440-449 [FREE Full text] [doi: [10.1111/j.1464-5491.2010.03222.x](https://doi.org/10.1111/j.1464-5491.2010.03222.x)] [Medline: [21204961](https://pubmed.ncbi.nlm.nih.gov/21204961/)]
54. Inamdar A. Correlation between fasting heart rate and fasting plasma glucose level in rural Indians. *Eur Heart J* 2022 Feb 04;43(Suppl 1):ehab849.158 [FREE Full text] [doi: [10.1093/eurheartj/ehab849.158](https://doi.org/10.1093/eurheartj/ehab849.158)]
55. Gupta S, Gupta RK, Kulshrestha M, Chaudhary RR. Evaluation of ECG abnormalities in patients with asymptomatic type 2 diabetes mellitus. *J Clin Diagn Res* 2017 Apr;11(4):OC39-OC41 [FREE Full text] [doi: [10.7860/JCDR/2017/24882.9740](https://doi.org/10.7860/JCDR/2017/24882.9740)] [Medline: [28571189](https://pubmed.ncbi.nlm.nih.gov/28571189/)]
56. Nirala N, Periyasamy R, Singh BK, Kumar A. Detection of type-2 diabetes using characteristics of toe photoplethysmogram by applying support vector machine. *Biocybern Biomed Eng* 2019 Jan;39(1):38-51 [FREE Full text] [doi: [10.1016/j.bbe.2018.09.007](https://doi.org/10.1016/j.bbe.2018.09.007)]
57. Philip LA, Rajasekaran K, Jothi ES. Continuous monitoring of blood glucose using photoplethysmograph signal. In: *Proceedings of the 2017 International Conference on Innovations in Electrical, Electronics, Instrumentation and Media Technology*. 2017 Presented at: ICEEIMT '17; February 3-4, 2017; Coimbatore, India p. 187-191 URL: <https://ieeexplore.ieee.org/document/8116832> [doi: [10.1109/icieeimt.2017.8116832](https://doi.org/10.1109/icieeimt.2017.8116832)]
58. Moraes JL, Rocha MX, Vasconcelos GG, Vasconcelos Filho JE, de Albuquerque VH, Alexandria AR. Advances in photoplethysmography signal analysis for biomedical applications. *Sensors (Basel)* 2018 Jun 09;18(6):1894 [FREE Full text] [doi: [10.3390/s18061894](https://doi.org/10.3390/s18061894)] [Medline: [29890749](https://pubmed.ncbi.nlm.nih.gov/29890749/)]
59. Habbu S, Dale M, Ghongade R. Estimation of blood glucose by non-invasive method using photoplethysmography. *Sādhanā* 2019 Mar 16;44(6):135 [FREE Full text] [doi: [10.1007/s12046-019-1118-9](https://doi.org/10.1007/s12046-019-1118-9)]
60. Hina A, Nadeem H, Saadeh W. A single LED photoplethysmography-based noninvasive glucose monitoring prototype system. In: *Proceedings of the 2019 IEEE International Symposium on Circuits and Systems*. 2019 Presented at: ISCAS '19; May 26-29, 2019; Sapporo, Japan p. 1-5 URL: <https://ieeexplore.ieee.org/document/8702747> [doi: [10.1109/iscas.2019.8702747](https://doi.org/10.1109/iscas.2019.8702747)]
61. Richman JS, Moorman JR. Physiological time-series analysis using approximate entropy and sample entropy. *Am J Physiol Heart Circ Physiol* 2000 Jun;278(6):H2039-H2049 [FREE Full text] [doi: [10.1152/ajpheart.2000.278.6.H2039](https://doi.org/10.1152/ajpheart.2000.278.6.H2039)] [Medline: [10843903](https://pubmed.ncbi.nlm.nih.gov/10843903/)]
62. Delgado-Bonal A, Marshak A. Approximate entropy and sample entropy: a comprehensive tutorial. *Entropy (Basel)* 2019 May 28;21(6):541 [FREE Full text] [doi: [10.3390/e21060541](https://doi.org/10.3390/e21060541)] [Medline: [33267255](https://pubmed.ncbi.nlm.nih.gov/33267255/)]
63. Nordström A, Hadrévi J, Olsson T, Franks PW, Nordström P. Higher prevalence of type 2 diabetes in men than in women is associated with differences in visceral fat mass. *J Clin Endocrinol Metab* 2016 Oct;101(10):3740-3746 [FREE Full text] [doi: [10.1210/jc.2016-1915](https://doi.org/10.1210/jc.2016-1915)] [Medline: [27490920](https://pubmed.ncbi.nlm.nih.gov/27490920/)]
64. Shi B, Yen SC, Tay A, Tan DM, Chia NS, Au WL. Convolutional neural network for freezing of gait detection leveraging the continuous wavelet transform on lower extremities wearable sensors data. In: *Proceedings of the 42nd Annual International Conference of the IEEE Engineering in Medicine & Biology Society*. 2020 Presented at: EMBC '20; July 20-24, 2020; Montreal, QC p. 5410-5415 URL: <https://ieeexplore.ieee.org/document/9175687> [doi: [10.1109/embc44109.2020.9175687](https://doi.org/10.1109/embc44109.2020.9175687)]
65. Shi B, Tay A, Au WL, Tan DM, Chia NS, Yen SC. Detection of freezing of gait using convolutional neural networks and data from lower limb motion sensors. *IEEE Trans Biomed Eng* 2022 Jul;69(7):2256-2267 [FREE Full text] [doi: [10.1109/TBME.2022.3140258](https://doi.org/10.1109/TBME.2022.3140258)] [Medline: [34986092](https://pubmed.ncbi.nlm.nih.gov/34986092/)]
66. Alwosheel A, van Cranenburgh S, Chorus CG. Is your dataset big enough? Sample size requirements when using artificial neural networks for discrete choice analysis. *J Choice Model* 2018 Sep;28:167-182 [FREE Full text] [doi: [10.1016/j.jocm.2018.07.002](https://doi.org/10.1016/j.jocm.2018.07.002)]
67. Lyssenko V, Jonsson A, Almgren P, Pulizzi N, Isomaa B, Tuomi T, et al. Clinical risk factors, DNA variants, and the development of type 2 diabetes. *N Engl J Med* 2008 Nov 20;359(21):2220-2232 [FREE Full text] [doi: [10.1056/NEJMoa0801869](https://doi.org/10.1056/NEJMoa0801869)] [Medline: [19020324](https://pubmed.ncbi.nlm.nih.gov/19020324/)]
68. Diabetes risk factors. US Centers for Disease Control and Prevention. 2022. URL: <https://www.cdc.gov/diabetes/basics/risk-factors.html> [accessed 2023-11-04]
69. Lundberg SM, Lee SI. A unified approach to interpreting model predictions. In: *Proceedings of the 31st Conference on Neural Information Processing Systems*. 2017 Presented at: NIPS '17; Dec 4-7, 2017; Long Beach, CA p. 4768-4777 URL: https://proceedings.neurips.cc/paper_files/paper/2017/file/8a20a8621978632d76c43dfd28b67767-Paper.pdf

70. Zhou XH, McClish DK, Obuchowski NA. *Statistical Methods in Diagnostic Medicine*. Hoboken, NJ: John Wiley & Sons; 2009.

Abbreviations

AI: artificial intelligence
BGEM: blood glucose evaluation and monitoring
CWT: continuous wavelet transform
DM: diabetes mellitus
G-mean: geometric mean
HR: heart rate
HRV: heart rate variability
IFG: impaired fasting glycemia
IGT: impaired glucose tolerance
KTE: Kaiser-Teager energy
ML: machine learning
MSE: multiscale entropy
PNS: parasympathetic nervous system
PPG: photoplethysmography
PSD: power spectral density
SampEn: sample entropy
SHAP: Shapley additive explanations
SNS: sympathetic nervous system
SVM: support vector machine
T1DM: type 1 diabetes mellitus
T2DM: type 2 diabetes mellitus

Edited by C Xiao; submitted 22.04.23; peer-reviewed by N Jiwani, S Tedesco; comments to author 04.07.23; revised version received 31.08.23; accepted 28.09.23; published 27.10.23.

Please cite as:

Shi B, Dhaliwal SS, Soo M, Chan C, Wong J, Lam NWC, Zhou E, Paitimusa V, Loke KY, Chin J, Chua MT, Liaw KCS, Lim AWH, Insyirah FF, Yen SC, Tay A, Ang SB

Assessing Elevated Blood Glucose Levels Through Blood Glucose Evaluation and Monitoring Using Machine Learning and Wearable Photoplethysmography Sensors: Algorithm Development and Validation

JMIR AI 2023;2:e48340

URL: <https://ai.jmir.org/2023/1/e48340>

doi: [10.2196/48340](https://doi.org/10.2196/48340)

PMID: [38875549](https://pubmed.ncbi.nlm.nih.gov/38875549/)

©Bohan Shi, Satvinder Singh Dhaliwal, Marcus Soo, Cheri Chan, Jocelin Wong, Natalie W C Lam, Entong Zhou, Vivien Paitimusa, Kum Yin Loke, Joel Chin, Mei Tuan Chua, Kathy Chiew Suan Liaw, Amos W H Lim, Fadel Fatin Insyirah, Shih-Cheng Yen, Arthur Tay, Seng Bin Ang. Originally published in JMIR AI (<https://ai.jmir.org>), 27.10.2023. This is an open-access article distributed under the terms of the Creative Commons Attribution License (<https://creativecommons.org/licenses/by/4.0/>), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR AI, is properly cited. The complete bibliographic information, a link to the original publication on <https://www.ai.jmir.org/>, as well as this copyright and license information must be included.

Original Paper

Insights on the Current State and Future Outlook of AI in Health Care: Expert Interview Study

Pia Hummelsberger¹, BSc; Timo K Koch^{1,2}, PhD; Sabrina Rauh¹, MSc; Julia Dorn¹, MSc; Eva Lermer^{1,3}, PhD; Martina Raue⁴, PhD; Matthias F C Hudecek⁵, PhD; Andreas Schicho⁶, MD; Errol Colak^{7,8,9}, MD; Marzyeh Ghassemi^{10,11}, PhD; Susanne Gaube¹², PhD

¹LMU Center for Leadership and People Management, Department of Psychology, LMU Munich, Munich, Germany

²Department of Psychology, LMU Munich, Munich, Germany

³Department of Business Psychology, Technical University of Applied Sciences Augsburg, Augsburg, Germany

⁴MIT AgeLab, Massachusetts Institute of Technology, Cambridge, MA, United States

⁵Department of Experimental Psychology, University of Regensburg, Regensburg, Germany

⁶Department of Radiology, University Hospital Regensburg, Regensburg, Germany

⁷Li Ka Shing Knowledge Institute, St. Michael's Hospital, Unity Health Toronto, Toronto, ON, Canada

⁸Department of Medical Imaging, St. Michael's Hospital, Unity Health Toronto, Toronto, ON, Canada

⁹Department of Medical Imaging, Faculty of Medicine, University of Toronto, Toronto, ON, Canada

¹⁰Electrical Engineering and Computer Science, Institute for Medical Engineering and Science, Massachusetts Institute of Technology, Cambridge, MA, United States

¹¹Vector Institute, Toronto, ON, Canada

¹²UCL Global Business School for Health, University College London, London, United Kingdom

Corresponding Author:

Pia Hummelsberger, BSc

LMU Center for Leadership and People Management

Department of Psychology

LMU Munich

Geschwister-Scholl-Platz 1

Munich, 80539

Germany

Phone: 49 89 2180 9773

Email: P.Hummelsberger@psy.lmu.de

Abstract

Background: Artificial intelligence (AI) is often promoted as a potential solution for many challenges health care systems face worldwide. However, its implementation in clinical practice lags behind its technological development.

Objective: This study aims to gain insights into the current state and prospects of AI technology from the stakeholders most directly involved in its adoption in the health care sector whose perspectives have received limited attention in research to date.

Methods: For this purpose, the perspectives of AI researchers and health care IT professionals in North America and Western Europe were collected and compared for profession-specific and regional differences. In this preregistered, mixed methods, cross-sectional study, 23 experts were interviewed using a semistructured guide. Data from the interviews were analyzed using deductive and inductive qualitative methods for the thematic analysis along with topic modeling to identify latent topics.

Results: Through our thematic analysis, four major categories emerged: (1) the current state of AI systems in health care, (2) the criteria and requirements for implementing AI systems in health care, (3) the challenges in implementing AI systems in health care, and (4) the prospects of the technology. Experts discussed the capabilities and limitations of current AI systems in health care in addition to their prevalence and regional differences. Several criteria and requirements deemed necessary for the successful implementation of AI systems were identified, including the technology's performance and security, smooth system integration and human-AI interaction, costs, stakeholder involvement, and employee training. However, regulatory, logistical, and technical issues were identified as the most critical barriers to an effective technology implementation process. In the future, our experts predicted both various threats and many opportunities related to AI technology in the health care sector.

Conclusions: Our work provides new insights into the current state, criteria, challenges, and outlook for implementing AI technology in health care from the perspective of AI researchers and IT professionals in North America and Western Europe. For the full potential of AI-enabled technologies to be exploited and for them to contribute to solving current health care challenges, critical implementation criteria must be met, and all groups involved in the process must work together.

(*JMIR AI* 2023;2:e47353) doi:[10.2196/47353](https://doi.org/10.2196/47353)

KEYWORDS

artificial intelligence; AI; machine learning; health care; digital health technology; technology implementation; expert interviews; mixed methods; topic modeling

Introduction

Background

Rising life expectancy, increasing prevalence of noncommunicable diseases (eg, diabetes), and staffing shortages are among the most severe challenges health care systems face worldwide [1]. As a result, the demand for health care services is steadily increasing, and health care costs are soaring [2,3]. Moreover, the high demand for services, extensive administrative and documentation requirements, and staffing shortages lead to heavy workloads for health care workers (HCWs) and reduce the time that staff can spend with patients and performing actual medical duties [4]. These circumstances jeopardize patient safety and limit the overall ability to deliver health care services [5-8].

The use of health technologies has often been suggested as a possible solution to address these challenges. By improving workflows, relieving staff of routine tasks, and reducing the frequency of medication errors and medical errors in general [8], health technologies might help ensure better health outcomes and increase efficiency [9,10]. In particular, artificial intelligence (AI) through machine learning has increasingly become the focus of health IT development in recent years. Health care professionals and patients associate AI technology with improved care [11,12] and reduced workloads [11,13,14]. Numerous high-performing AI algorithms have been developed to support HCWs with various tasks in different medical fields, such as radiology, cardiology, neurology, ophthalmology, oncology, gastroenterology, mental health, and many others [15-17].

However, despite the extensive research on AI applications in health care, the implementation of AI-enabled clinical decision support systems (AI-CDSSs) in clinical practice lags behind what would be feasible according to technical developments [18]. Several explanations for the slow adoption of AI systems in health care have already been proposed. Various groups are involved in this AI implementation process: (1) policy makers and authorities who determine the framework conditions for the entire process; (2) researchers and developers who develop, train, and market the system with its various functions; and (3) IT experts in health care facilities who sometimes make decisions about system acquisitions, integrate them into the existing infrastructure and maintain them if necessary, and introduce them to the (4) HCWs who ultimately use the system in their everyday work [19,20]. Many issues have been brought forward by or are attributed to 2 groups of people on both ends of the technology implementation spectrum: HCWs and policy

makers, both of whom are essential for the success of AI technologies in health care.

On one side of the spectrum, physicians and other HCWs are the end users of most AI systems in health care. The technology is developed to support their workflows, but if HCWs are reluctant to use AI systems, the proposed advantages of the technology cannot materialize [21,22]. On the one hand, HCWs believe that AI has the potential to improve the quality of care through more accurate and precise diagnoses, as well as enabling faster diagnoses and shorter wait times. It can also promote personalized care tailored to the patient and ensure greater consistency in diagnoses as the performance of AI technologies does not suffer from human stress symptoms, fatigue, or difficulty concentrating [23,24]. HCWs also expect collaboration with AI-enabled systems to reduce daily workload and save staff time by allowing the technology to prioritize symptoms and patients and provide legal protection for medical staff through ongoing documentation of the care process [23,24]. On the other hand, however, research has shown that current and future HCWs are reluctant to use AI applications in their daily work for a variety of reasons. These include concerns about the performance of the technology and fears that overtechnologization may impair their abilities over time as AI takes over tasks and clinicians become overly reliant and accustomed to the technology [13,23]. Some HCWs also suspect that AI systems will influence staff diagnostic decisions [23] and fear that technology will make their jobs redundant [25,26]. In addition, HCWs are concerned that using these technologies will negatively affect the physician-patient relationship and might compromise privacy as AI systems would have to work with patients' sensitive data [23].

On the other side of the spectrum are policy makers (eg, intragovernmental and governmental organizations as well as regulatory bodies such as the US Food and Drug Administration [FDA] or the European Medicines Agency). They are responsible for the ethical, legal, and regulatory frameworks and conditions for the implementation of AI systems in health care. Policy-making bodies have already issued guidelines on AI implementation and have discussed unresolved legal and regulatory issues such as certification [27], liability, and data protection [28]. Moreover, policy makers have expressed concern about ethical issues such as discrimination and lack of transparency, which might hinder the safe and widespread implementation of AI applications in health care [27].

However, when it comes to the physical implementation of AI technology into the existing health care infrastructure, in most cases, neither policy makers nor HCWs are directly involved

in the process. In reality, the 2 other stakeholder groups (ie, researchers and developers as well as IT experts) are responsible for the practical implementation of AI products in health care facilities. Researchers have discussed many challenges surrounding AI systems in health care. Some of these are naturally linked to the issues raised by the other stakeholder groups, such as the lack of trust among users [21,29]; regulatory burdens; and concerns about accountability, ethical data use, biases, and discrimination [30]. Other mentioned challenges relate to more technical issues such as unsatisfactory system performance, detection of biased data, system explainability, cost and quality of labeled data, and computational limitations [30,31]. The perspective of IT professionals has received considerably less attention in the literature. Some research has shown that they see the lack of human, professional, and financial resources and incompatibility with existing IT infrastructures as barriers to implementing AI technologies in health care [32]. In addition to an acceptable user interface and robust connectivity to the infrastructure, AI researchers and developers are searching for contact with clinical users [33].

Besides looking at the various stakeholder groups, it is important to consider regional differences when trying to obtain a global perspective on the current state of AI implementation in health care. International comparisons show substantial differences in overall investment in developing and deploying new AI technologies. Overall, the United States and China have raised the most venture capital funds, followed by Europe, which, however, lags significantly behind the former 2 [34]. When looking specifically at health care–related investment in AI technology, again, the United States, China, and Europe are the 3 global players, which is also reflected by their amount of research output [35]. One study has already conducted a cross-regional comparison of the adoption of AI in small- and medium-sized health enterprises between Germany and China. It showed that Germany-based professionals named challenges related to data accessibility, transparency, and regulations more often than their Chinese colleagues [36]. To the best of our knowledge, no study has systematically compared European and North American experts' views on the opportunities and challenges of implementing AI applications in health care.

Objectives

This study focused only on the 2 professional groups closest to the physical integration of AI systems. We wanted to mention that the topic should ideally be viewed more holistically. According to the Responsible Innovation and Responsible Research and Innovation approaches, it is important to involve all stakeholders to prioritize the ethical, social, and sustainable aspects of technological advances. This is to ensure that innovation and research benefit society while minimizing harm and accounting for societal needs and values [37].

By exploring the implementation of AI technologies in health care, we wanted to focus on stakeholders directly involved in the process. Researchers' perspectives have been discussed extensively in the literature but have mainly focused on potential opportunities, technical challenges, and ethical issues of AI models rather than their implementation. In contrast, the views of IT professionals in health care have received little attention

overall. To fill this gap, this study used a mixed methods approach to collect and compare the opinions of researchers and IT professionals on implementing AI technology in health care from their respective points of view. In addition, we included respondents from North America and Europe to uncover potential regional differences in addition to profession-specific differences.

Methods

Sample

The 2 critical inclusion criteria for participating in this preregistered study were *profession* and *region*. We focused on researchers working on AI applications for health care and medicine and IT professionals in the health care sector. These professional groups allowed us to obtain the views and differences in opinions of 2 key stakeholders directly involved in the implementation of AI applications in health care practice. The researcher group consisted of computer scientists and clinical scientists ranging from senior doctoral candidates to faculty members. The group of IT experts included chief technical officers and chief information officers from hospitals, representatives of medical device safety organizations, and chief executive officers of health IT companies. The 2 regions of interest were Western Europe and North America, with the European countries of Germany, Austria, Switzerland, and Belgium and the North American countries of the United States and Canada being represented. By including participants from these Western regions, we were able to gain valuable insights into different legal and health care systems and highlight regional differences between these global players. As the 2 professional groups are highly specific, no other selection criteria or prerequisites, such as minimum professional experience, were stipulated. Ultimately, 23 individuals were interviewed, including 13 (57%) researchers (n=7, 54% from Western Europe and n=6, 46% from North America) and 10 (43%) IT experts (n=8, 80% from Western Europe and n=2, 20% from North America).

Recruitment

Sampling was performed via a web search based on relevant publications and matching of LinkedIn profiles. In addition, experts were recruited via snowball sampling through the authors' networks and recommendations from participants and other third parties. The participants were selected on a nonprobabilistic basis, that is, deliberately according to the aforementioned criteria [38]. We planned to interview at least 20 experts, balanced between professional groups and regions, to obtain a well-rounded picture of the topic. A total of 104 candidates were approached via email, of whom 23 (22% participation rate) agreed to participate. Interviewees received no compensation for their participation.

Data Collection

Data for this cross-sectional, mixed methods study were collected using semistructured expert interviews. Consequently, all participants received the same questions from the interview guide, with the option of the interviewers asking follow-up questions or using prompts if needed. The authors developed

the interview guide for this study based on the research questions and the literature presented in the *Introduction* section. It included questions from four categories: (1) the prevalence of AI applications in hospitals and the current state of the technology, (2) their implementation criteria, (3) the challenges, and (4) the potential of implementing AI systems in health care. The original interview guide was pretested twice, resulting in minor improvements. Between November 2021 and January 2022, all 23 interviews were conducted remotely via Zoom (version 5.8.3-5.9.1; Zoom Video Communications, Inc) and by phone. The interviews lasted between 14.5 and 49.5 (mean 30.0, SD 8.0) minutes and were conducted in English (19/23, 83%) and, at the request of the interviewees, in German (4/23, 17%). At the start, participants were informed about the purpose, procedure, expected duration, voluntary nature of the interview, and how their data would be processed. The interviewees provided informed consent to participate in the study and for the interviews to be recorded. At the beginning of the recording, the participants were first asked to briefly describe their professional backgrounds as an icebreaker. This was followed by our predefined interview questions and, if needed, follow-up questions and prompts. After discussing all the questions, participants had the opportunity to add anything they felt was relevant to the topic. At the end of the interview, we asked for other potential interviewees and thanked the interviewees for their participation.

Data Preparation

Every participant received a nonidentifiable acronym under which their materials were stored and analyzed. The acronym only indicated the person's professional group and region, which was needed for the analysis (researcher in North America [RENA], researcher in Western Europe [REEU], IT expert in North America [ITNA], and IT expert in Western Europe [ITEU]). The interview recordings were transcribed using Trint (version unknown; Trint Limited). Contextual information that could lead to the identification of an individual was manually anonymized in the transcripts. All transcripts were reviewed, translated into English if necessary, and uploaded to MAXQDA (version 20.4.2; VERBI Software GmbH). The raw material with sensitive data, that is, consent forms and audio or video files, was securely stored in a password-protected digital folder.

Data Analysis

For thematic analysis of the data, we used MAXQDA. We chose a combination of deductive and inductive qualitative methods [39]. This approach integrates a theory-driven template [40] and a data-driven framework [41] for developing codes. The method includes 6 steps for data analysis, the details of which can be found in the literature [39] and in an extra document in the study's repository on the Open Science Framework [42]. At the end of the thematic analysis process, 14 cross-cutting themes and 172 subthemes were identified, divided into 4 categories, and captured in the final codebook. Three additional themes were identified in the *challenges* category: interdisciplinary work, ethics, and user. However, these were much smaller in scope than the other themes and, therefore, were not considered further in the rest of the study. To validate the coding process, a third previously uninvolved author

analyzed a representative subsample of 10% (10/100) of the data using the final coding manual [43,44]. The intercoder agreement with a code overlap in segments of at least 90% (90/100) was a Cohen κ value of 0.77, which is considered a substantial match [45,46]. After the second coding, only minor changes were made to the final codebook.

Following the qualitative thematic analysis, we also analyzed the interviewees' responses quantitatively using topic modeling to identify latent topics as well as the most frequently used words. Quantitative text analysis has been found to be a useful tool for validating results of previous qualitative analysis [47-49]. In this case, we first removed the introductory and closing parts of the interviews that only contained introductions and small talk. Furthermore, we deleted all stop words, which are words that are commonly used with little or no relevance to the content of a text (eg, "and" and "did"). We also singularized all words (ie, "algorithms" became "algorithm"). Then, we computed the frequencies of words (uni-, bi-, and trigrams) grouped by interviewees' region and profession. For a better visual illustration, these were plotted in word clouds. On the basis of the findings from the qualitative analyses that had been validated by quantitative analysis, we extracted 14 topics using latent Dirichlet allocation [50] with Gibbs sampling (Cronbach $\alpha=0.30$) [51]. Finally, we manually matched the qualitative themes with the quantitatively extracted topics with regard to their content. All text data processing and statistical analyses were performed using the statistical software R (version 4.1.1; R Foundation for Statistical Computing). Specifically, we used the R packages *udpipe* for tokenization [52] and *topicmodels* as well as *ldatuning* for topic modeling [49,53].

The following documents can be found in the Open Science Framework repository [42]: the preregistration, the list of participants, the interview instructions and the interview guide in English and German, the description of the qualitative coding process, the final codebook, a table showing the frequency of themes and subthemes, and a table of the top 10 words identified during quantitative topic modeling.

Ethical Considerations

This study was exempt from a full ethical review by Committee on the Use of Humans as Experimental Subjects, the institutional review board of the Massachusetts Institute of Technology, by meeting the criteria for exemption (E-4248).

Results

Overview

Four broad categories emerged during the qualitative analysis: (1) the current state of AI technology in health care, (2) the implementation criteria and requirements of AI systems in health care, (3) the challenges in implementing AI technology in health care, and (4) the technology's outlook. [Table 1](#) provides an overview of the most relevant aspects that emerged from the qualitative (themes) and quantitative (topics) analyses clustered within these 4 categories. [Table 2](#) provides a more detailed overview of the 14 topics that emerged from the quantitative analysis, each with the top 5 underlying words. Initially, we present a quantitative overview of the interview content of each

expert group. This is followed by an in-depth look at the most relevant qualitative themes. Subthemes that fall under several themes are described only once.

The 14 topics extracted through the quantitative analysis of our interview data matched well with many themes from the qualitative analysis: prevalence (topics 1 and 7), regional

differences (topic 2), capabilities (topics 3 and 7), limitations (topic 5), performance and safety (topics 5 and 9), system integration and human-AI interaction (topics 8 and 12), costs (topic 11), stakeholder involvement (topic 11), employee training (topic 4), different kinds of challenges (topics 3, 11, 12, and 13), threats (topic 11), and opportunities (topics 6, 10, and 14).

Table 1. Relevant themes and topics from the qualitative and quantitative analyses.

Category and theme (qualitative)	Topic (quantitative)
Current state of AI^a systems in health care	
Prevalence	<ul style="list-style-type: none"> • AI in health care • AI in medical imaging^b
Regional differences	<ul style="list-style-type: none"> • Regional challenges^b
Capabilities	<ul style="list-style-type: none"> • Improving the everyday experience of HCWs^c • AI in medical imaging^b
Limitations	<ul style="list-style-type: none"> • Clinical research^b
Implementation criteria and requirements of AI systems in health care	
Performance and safety	<ul style="list-style-type: none"> • Clinical research^b • Performance
System integration and human-AI interaction	<ul style="list-style-type: none"> • Workflow optimization • Human-AI interaction^b
Costs	<ul style="list-style-type: none"> • Barriers to AI implementation^b
Stakeholder involvement	<ul style="list-style-type: none"> • Barriers to AI implementation^b
Employee training	<ul style="list-style-type: none"> • Employee training
Challenges in implementing AI systems in health care	
Regulatory challenges	<ul style="list-style-type: none"> • Regional challenges^b
Logistical challenges	<ul style="list-style-type: none"> • Barriers to AI implementation^b
Technical challenges	<ul style="list-style-type: none"> • Human-AI interaction^b • Industry challenges
Outlook	
Threats	<ul style="list-style-type: none"> • Barriers to AI implementation^b
Opportunities	<ul style="list-style-type: none"> • Future developments • Technical advances • Opportunities

^aAI: artificial intelligence.

^bThese quantitative topics can be assigned to several qualitative themes.

^cHCW: health care worker.

Table 2. A total of 14 quantitatively extracted topics from the interview transcripts.

Topic name	Topic description	5 most frequent words ^a
AI ^b in health care	Impact of AI adoption on health outcomes	health care, perspective, health, challenge, learn
Regional challenges	Regulatory challenges for AI implementation in certain regions	germany, company, clinic, regulatory, country
Improving the everyday experience of HCWs ^c	Integrating AI into routine patient care for the benefit of HCWs	patient, time, nurse, care, day
Employee training	Training HCWs on how the system works and its limitations	system, decision, implement, training, medical
Clinical research	Studying the risks and benefits of AI in clinical settings	algorithm, physician, clinician, study, risk
Future developments	Exploring long-term solutions and regulatory aspects for diagnostic development	solution, term, future stuff, united states
AI in medical imaging	Current AI applications in radiology and medical imaging	image, radiologist, radiology, diagnosis, application
Workflow optimization	Improving institutional workflows and communications with AI	model, question, understand, talking, sense
Performance	Impact of AI on performance and practices	human, data, performance, practice, super
Technical advances	Using technology to facilitate knowledge-based change in medicine	technology, field simply, machine learning, change
Barriers to AI implementation	Logistical and stakeholder challenges in implementing AI	person, situation cost, university, feel
Human-AI interaction	User-centered technology integration to support HCWs	hospital, doctor, environment, person, issue
Industry challenges	Industry challenges in deploying AI systems	process, level, wrong, set, improve
Opportunities	Creating opportunities for data-driven clinical care in specific domains	data, clinical, basically, care, answer

^aThe 10 most frequent words are included in the project repository.

^bAI: artificial intelligence.

^cHCW: health care worker.

Quantitative Overview of Regions and Professions

Table 3 shows the 10 most frequently used words divided by professional (researchers and IT experts) and regional groups

(Western Europe and North America). The higher the words are ranked in the table, the more frequently they were mentioned by the interviewed experts.

Table 3. Ten most frequently used words grouped by profession and region of the interviewees.

Profession	Western Europe	North America
Researchers	<ol style="list-style-type: none"> 1. person 2. data 3. patient 4. system 5. algorithm 6. doctor 7. germany 8. hospital 9. process 10. human 	<ol style="list-style-type: none"> 1. algorithm 2. person 3. data 4. patient 5. hospital 6. healthcare 7. human 8. time 9. care 10. model
IT experts	<ol style="list-style-type: none"> 1. patient 2. algorithm 3. data 4. hospital 5. company 6. person 7. germany 8. question 9. time 10. diagnosis 	<ol style="list-style-type: none"> 1. technology 2. clinical 3. health 4. adoption 5. clinic 6. level 7. model 8. opportunity 9. augmentation 10. challenge

Current State of AI Systems in Health Care

Prevalence of AI Systems in Use

Respondents named 12 different medical fields or specialties in which AI algorithms have been developed for clinical practice: neurology, oncology, radiology, dermatology, cytomorphology, surgery, pediatrics, pathology, ophthalmology, urology, genomics, and diabetology, as well as intensive care medicine. Many interviewees focused on AI systems for radiology, which could indicate that this is the most mature field for the technology. Classification of medical imaging was often mentioned as a relevant use case, again potentially highlighting the maturity of this application. Interviewees also mentioned discipline-independent use cases. For instance, current AI systems can also use text-based data from electronic health records (EHRs) to make medical predictions using natural language processing. AI algorithms are also used to optimize administrative tasks such as staff scheduling and billing. However, 48% (11/23) of the interviewees acknowledged that many systems are not commercially available but are at the stage of in-house scientific research projects. Almost exclusively, European interviewees emphasized that systems are not widely used in routine clinical practice:

We see projects on the scientific side where we use AI. But I couldn't describe a single use case where a real AI, some kind of neural network deep learning mechanism, would be in place in our normal health care activities. [ITEU18; position 7]

Regional Differences in Research and Development

The interviewees mentioned the United States and China as leaders in AI research, whereas Germany and many European countries seemed to lag behind. Within Europe, the Nordic and Baltic countries, as well as the United Kingdom, are considered frontrunners in AI development for the health care sector:

So if you look at places like Singapore and also China, you will also see that this area of [sic] analyzing huge amounts of data and applying algorithms from AI [sic] to novel case [sic], this is something where they are, I would say, even years ahead of what we [Germany] are doing. [ITEU18; position 11]

Several reasons were given for why European AI research is trailing that of the United States and China. Researchers and IT experts primarily blamed the lack of available data for training the models caused by stricter data protection laws and regulations. The General Data Protection Regulation (GDPR) implemented in the European Union in 2018 makes sharing data between research and health care facilities within and across countries more complicated:

I think that GDPR...makes it a little more difficult for data sharing in Europe. And so that may be part of why the research is not...progressing quite as fast. [RENA05; position 20]

Researchers from Europe further pointed to the slow progress of digitalization in health care and the lack of financial

investments as barriers to the advancement of AI-enabled systems:

Germany is lagging behind due to digitalization...a switch from spreadsheets to platforms that really integrate patient data is really needed in Germany. [REEU10; position 13]

Capabilities of Current AI Systems

Both professional groups referred to similar technical capabilities. Currently, AI algorithms can support HCWs mostly in 2 ways. First, the systems can perform specific, highly repetitive tasks that are easy but time-consuming for humans. Consequently, deploying these applications can reduce workload and free up time for other tasks:

I should say here...that the AI applications are usually very narrow based, which means that they can do a simple task...But it's automated, so it might go faster, which is easier for the radiologist. [ITEU08; position 48]

Second, AI systems outperform humans when working with large and complex data. This type of data is often characterized by a diffuse structure, complex interrelationships, and multidimensionality. By instantly incorporating more data than a human ever could, AI algorithms can make faster and more accurate predictions:

The way these algorithms work is they can handle...complexity that we as humans can't. [RENA16; position 33]

Limitations of Current AI Systems

Both researchers and IT professionals described the fact that AI algorithms currently cannot operate without human supervision as a major limitation. At the moment, it is required that HCWs verify the algorithms' results. Thus, the full responsibility and liability for clinical decision-making remain with the user:

The limits, obviously, are [sic] they can't take responsibility for what they're doing...They can't take any responsibility in terms of medical legal issues. So whenever you do something...some poor doctor has to sign the whole thing and then he's responsible for whatever happens. [REEU07; position 28]

Another limiting factor mentioned by both groups was the technology's high task specificity, which limits its usefulness in 2 ways. On the one hand, AI algorithms are often programmed to use only one source of information (eg, 1 type of medical image) for their prediction, whereas integrating multiple sources (eg, medical images and patient history) would yield better results. On the other hand, medical decisions often require the involvement of multiple disciplines, for example, radiology and surgery. Consequently, integrating several stand-alone algorithms or developing multitask algorithms would be needed to support the entire workflow for multidisciplinary teams:

That's just that...the AI system is trained for a specific use case, for example skin cancer, then it looks at the skin image, but does not include other things, from

the case history or similar large. [ITEU09; position 24]

One fundamental limitation mentioned by IT professionals was the lack of explainability of currently deployed AI algorithms. The absence of information on how the algorithm operates makes it difficult for users to understand why a specific recommendation or prediction was made, which might make them skeptical of relying on it:

So it's always the case that they say the systems are great, but mostly they can't explain them reasonably. That means that one of the current limits is the ability to explain how the decision was actually made. [ITEU09; position 23]

Implementation Criteria and Requirements for AI Systems in Health Care

Performance and Safety

Both professional groups mentioned high performance in the form of a low error rate most frequently as the primary criterion for adopting AI-enabled systems. Accordingly, algorithms should only be implemented in health care settings if they show high accuracy to ensure patient safety:

Because human lives are at stake here. Currently, there is simply no time for trivialities, but it must work 100.0%. And that's why over 99.0, so 99.5/99.8 are the requirements for implementing the AI system. [ITEU03; position 28]

Some interviewees advocated comparing the performance of algorithms with human performance and evaluating them using the same standards. However, in reality, users seem to have much higher performance expectations of AI systems than of humans. Therefore, the experts argued that algorithms with a performance that matches or exceeds that of human experts, even if they are not always perfect, might help improve overall decision accuracy:

If I have an algorithm whose AUC is .94. Really [sic] good performing algorithm. But the clinicians perform better. Their AUC is .96. It's not a good algorithm because you're not outperforming clinicians. But if you've got an algorithm where the performance isn't very good, their AUC is .68. But the clinician AUC is .58. It's a good algorithm because it does better than the clinicians. [RENA16; position 38]

Researchers emphasized that algorithms must be revalidated when deployed in a new environment as local data might differ from the training data. Therefore, all stakeholders involved in the implementation process must ensure that the algorithms perform well in new environments and over time:

Ideally, you would do a revalidation of that algorithm on your institution's data, your patient population. [RENA12; position 24]

System Integration and Human-AI Interaction

The experts pointed out that the successful deployment of AI systems depends on how easily they can be integrated into the existing technical infrastructure of the respective institution:

And how easy is it to integrate within the system? How much time does it take to do that? How many and how can we see the results? How can it be integrated in our reports, for example? [ITEU08; position 64]

Both professional groups considered good usability and smooth workflow integration to be as important as performance for using AI technology in health care. According to the experts, users will only be willing to engage with a new technology if it makes their work easier. System interfaces should be designed intuitively enough for users to operate without substantial training and must be adaptable to users' needs:

For the nurses, we actually had to develop an interface that they wanted to see. That's very simple for them to work with. So keeping things as simple as possible. [RENA16; position 48]

According to 35% (8/23) of the respondents, users' acceptance of and trust in the technology is another essential factor (or barrier if missing) before purchasing and deploying AI systems in health care settings. Without the end users' acceptance and willingness to use these systems, the implementation process is doomed to fail:

But [work] culture is way more important [than performance]. So culture first, do they actually want to use this stuff? Are they open-minded and they want to embrace that? [RENA16; position 37]

Several researchers even suggested that AI technology should work only in the background, automatically taking signals from all the different data streams, integrating them, and acting accordingly without human intervention. This would make the system much easier to use and bypass complex human-technology interaction issues. The interviewees claimed that background operability might be particularly advantageous regarding user acceptance as issues of trust in the technology might not even arise in solely background operating systems. In addition, less effort and fewer resources are needed to introduce the AI system to users if they are not directly interacting with it:

So when you look at what the future would hold, what's actually going to get adopted, I think they're going to be solutions that are doing operational things where the healthcare workers are not interacting in a deliberate way with those systems. [RENA15; position 29]

Although many experts argued that AI must be explainable so that users know how and why the system makes a prediction, some IT professionals completely disagreed. According to them, staff do not even need to know whether the underlying technology is AI-enabled so that they handle all devices unbiased:

Nobody should know there's an AI model inside. Is [sic] not relevant. [ITNA04; position 33]

Costs

Costs were also mentioned as a criterion by both professional groups. However, the interviewees disagreed on how important this factor is:

I would say, this is not the major factor. I mean, costs are always a factor, but in the end, it has to be evidence-based. In the end, you have to understand what is the outcome of using such an algorithm. [ITEU18; position 21]

I'm going to have to cough up a lot of money and then when am I going to see the value of this? So it's really important when it comes to the implementation that there is a very clear business case and value proposition of why this matters now, both near-term and long term. [ITNA02; position 26]

Stakeholder Involvement

There was only partial consensus on which actors are the most important for the implementation of AI systems in health care. This could be because technology procurement processes vary widely across health care facilities. Differences were also found between professional groups as well as between regions.

In both regions, the institution or department heads appear to be the driving force behind technology adoption. European researchers assumed that finance departments also play a role in purchasing decisions:

So in the end it's always the heads of the institutes or the chief physicians who have to say yes...So I would say that they are the ones who mainly have to be convinced. [ITEU09; position 32]

Researchers from North America stated that hospitals' IT professionals are involved in the implementation:

The other stakeholders are typically the people who manage the...computer systems and the people who would have to set it up and install it. [RENA12; position 31]

Interviewees from both professions and regions indicated that regulatory bodies are important stakeholders for implementing AI in health care:

If it's not built in the institution, you would have to go to actual regulatory approval. Certainly, if this system is going to have a direct impact on patient care. [ITNA04; position 22]

Moreover, some respondents also mentioned that the actual end users, meaning patients, might be a relevant stakeholder group for successfully implementing AI systems in health care:

And there again, we have the question: are we allowed to do so? Is it something the patient has to agree for and so on? So these are all criteria to choose. [ITEU18; position 21]

Employee Training

Nearly all experts emphasized the need for basic AI skills and knowledge so that users can safely interact with the systems and recognize their limitations. It has been argued that training

on AI should be integrated into the curricula for current and future HCWs who will work with AI-enhanced systems:

Part of the education of our workforce, will include the basics of how these systems work, where they fail, where they can potentially cause harm. [RENA15; position 31]

However, participants disagreed on how much training is needed. Some thought that HCWs need to be able to operate the AI systems and understand their underlying mechanisms, including functions and limitations. Consequently, training should start as early as possible, preferably already during the education period. Other experts thought that training should be limited to the most necessary information to minimize the burden on staff. In particular, the level of training should be adapted to the complexity of the AI system and the learning culture within institutions:

So what we are trying to do is to have students, first of all, be aware of artificial intelligence and what it is, what it can do, what it can't. Then different techniques like, for example, what is computer vision? How does that work? So what is object recognition? Then further on with natural language processing. [REEU10; position 48]

Challenges in Implementing AI Systems in Health Care

Regulatory Challenges

Data protection and security emerged as the primary regulatory challenges. Strict regulations limit access to data needed to develop advanced algorithms. Interviewees from Europe especially lamented that the inability to share data across institutions hinders AI research and implementation:

When you take machine learning...the regulatory challenges are the data protection regulation. [ITEU22; position 35]

Moreover, the experts mentioned that certification processes, especially FDA approval for medical products, are a significant challenge for developers. Documentation guidelines interfere with the continuous improvement of the algorithm once systems have been deployed:

Does he have the certificate? Has the constancy test been carried out? Every small deviation in patient monitoring must be documented, and this is also queried, sometimes half a year later, although the patient has long left. [ITEU03; position 47]

Predominantly, IT professionals were concerned about liability issues in cases when the system fails and incorrect decisions are made as a consequence:

And you can ask the question who's liable: the hospital or is [sic] the company that created the model? And we haven't seen the first lawsuit yet. [ITNA04; position 48]

Although regulations can slow down the development and implementation of AI systems in health care, some experts said that they are necessary to ensure patient safety:

Well, the regulatory process is inherently conservative...as slow as it needs to be to make sure that we stay safe and that's appropriate. [RENA05; position 48]

Logistical Challenges

Securing funding for AI algorithms was the most frequently cited logistical challenge. Both developers and medical institutions face high costs in developing, acquiring, implementing, and maintaining new systems:

The costs of healthcare are rising and rising in Germany and in other countries, too, so hospitals do not have all the money in the world to introduce the systems. [REEU06; position 45]

I know that the implementation is going rather slow, and for the vendors, it's slower than expected, which also makes it quite difficult for them because they have to invest a lot of money, and they have invested. But they also would like to see a return on investment, of course. [ITEU08; position 33]

The lack of IT professionals needed to implement AI-enabled systems into existing IT infrastructure was also mentioned as a huge barrier. In addition, in-house data scientists who can monitor and operate the systems are required, placing even greater human resources and financial burdens on institutions:

And this may include lack of access to IT resources and personnel, right, skilled people. [RENA15; position 18]

Some researchers pointed out that health care institutions need to collaborate more to improve AI algorithms and unlock their real potential. Collaboration mainly involves sharing and integrating data across institutions as, at the moment, important data for optimal predictions are lost for an algorithm when patients change institutions during their treatment. However, sharing and integrating sensitive data is particularly complex and resource intensive:

Healthcare is not a single point event. It's a process. And so somebody will go to his doctor and will get a potential diagnosis. We get some diagnostic workup. We'll go to the specialist, we'll get some more diagnostic work up. The information from the primary doctor gets lost. [REEU07; position 62]

Technical Challenges

Researchers identified the lack of available high-quality preprocessed training data and data on rare diseases as a major challenge affecting the algorithm's performance. In addition, using unprocessed hospital data (eg, data coming directly from EHRs), which would be more readily available, is challenging as these data are not standardized:

So you have label data and unlabeled data, and the labeled data is usually labeled by human experts. And the quality of the model always depends on whether or not the labels are accurate. [REEU10; position 18]

IT professionals were concerned about biases in the training data that could distort the algorithms and make their predictions less accurate for people who were underrepresented in the training data. Biases in the form of under- or overrepresentation of certain patient and disease groups can occur. For instance, wealthy and renowned hospitals, which are regularly involved in generating training data, have a nonrepresentative patient and disease pool. This is especially problematic as it is challenging to detect biases in the data in the first place and to correct the model at the operational level:

It's also very difficult to identify whether there is a certain bias involved. If you have a large set of data and we know that there are typically some biases and there is research to identify biases, but there's very often a hidden bias which you cannot automatically detect. [ITEU18; position 18]

Researchers also complained about the poor and inflexible IT infrastructure that makes the implementation of AI algorithms challenging:

And then when we speak about technical challenges, it's more about the hardware, to be honest, because although this is not always available in medical institutions. [REEU20; position 32]

Moreover, some interviewees mentioned that AI developers struggle to design AI system interfaces that meet user needs in the complex health care environment. Currently, the systems often fail to provide user-centered and user-friendly designs:

Then the other challenge is designing the human interaction in the [sic] way that people can actually use it. [REEU20; position 33]

Outlook

Threats

Researchers in particular expressed great concern about the possibility that the deployment of AI-enabled systems might exacerbate health care disparities that already exist in society. There are several reasons for this. As mentioned previously, biases in the algorithm's training data might lead to less accurate algorithmic predictions for underrepresented, often marginalized groups, which might cause serious harm. Moreover, health care facilities in wealthier regions tend to be the first to adopt new technologies. As a result, their patients will benefit from AI innovations, whereas patients in poorer areas will be left further behind:

There's a substantial risk for creating new or exacerbating existing racial, sexual and socioeconomic healthcare disparities. [RENA12; position 56]

Another threat mentioned only by researchers was automation bias, which is the tendency to rely too much on AI-CDSSs. As a result, system users may fail to detect prediction errors if they accept AI advice unconditionally. Consequently, automation bias poses a danger if the algorithm is not highly reliable, which could lead to many medical errors:

You can also have things go the other way where people put, you know, way too much trust in the AI, and they kind of, you know, blindly...trust whatever it's saying. Even...if they'd stopped and thought about it, they would realize that the result that was coming out is nonsensical. [RENA12; position 55]

IT professionals expressed concerns about cyberattacks as AI systems in health care are also not immune to hacking. Cyberattacks could affect both data security and patient outcomes if the algorithms are compromised or unavailable because of the attack:

We had some hacker attacks in the history, in the last 5 years in some hospitals in Europe and if systems are not available, then still all the work flows need to be working. And if you rely too much on AI and digitalization, of course, it's a problem. [REEU06; position 48]

Although it is often discussed that AI systems could make some jobs obsolete, our interviewees unanimously predicted that the adoption of AI technology will not lead to job losses in health care in the near future. However, task-specific skills that require a lot of training might decline if AI systems are widely used:

So whereas in the early days in the media, you could read AI will replace radiologists. Well, this is of course not true because looking at a CT scan of the lungs is much more than only counting nodules. [ITEU08; position 48]

If you have a system that supports you a lot, you may also have the risk to lose [sic] your own skill in a situation of doubt that can be very harmful. [REEU20; position 24]

Opportunities

According to the experts, the most significant opportunity for using AI systems in health care is the reduction in workload. For instance, outsourcing time-consuming and repetitive tasks to an AI system would allow HCWs to focus on more complex tasks and patient interactions:

When it's implemented in a very good way and the doctors have trust, it frees time for direct communication with the patient. [ITEU22; position 48]

IT experts saw tremendous opportunities in AI technology to improve diagnostic accuracy and patient outcomes through decision support. In addition, AI algorithms could enable truly personalized health care by analyzing multiple sources of health data simultaneously and across time. For instance, long-term EHRs could be combined with vital signs recorded via digital devices and analyzed using an algorithm. Long-term integrated data analysis could potentially facilitate the early detection of previously hidden disease patterns and provide individualized prevention and treatment plans:

So I do think that AI will be able to provide a more specific and more patient-specific treatment based upon the information, the data that we obtain. [ITEU08; position 75]

Researchers pointed out that health care logistics such as supply chain management and billing could benefit from AI systems. AI algorithms are already used in other industries to support logistical, administrative, and planning processes:

There's a lot of opportunity for AI in supply chain, billing, claims management. [ITNA04; position 44]

Discussion

Principal Findings and Comparison With Prior Work

Plenty of research on the challenges and opportunities of AI technology in health care has been published. However, our novel approach of pooling the expertise of AI researchers and IT professionals from Western Europe and North America resulted in a novel, nuanced, and comprehensive overview based on four main categories: (1) the current state, (2) implementation criteria and requirements, (3) implementation challenges, and (4) future outlook.

Within the current state theme, the interviewees mentioned that AI systems have been developed for various medical fields and use cases, primarily image classification in radiology and pathology, but have yet to be widely deployed in clinical practice. According to the literature, the 3 global players in health AI are the United States, China, and Europe, with the former 2 investing the most in research and development [34-36]. Our experts agreed that the United States and China dominate research and development but emphasized much more that Europe lags behind, largely because of lower investment in technology and digitalization and limited access to data because of stricter privacy regulations. At the moment, AI systems can support clinical decisions and diagnoses by providing predictions for specific tasks. Previous studies have found that many HCWs believe that AI systems will improve diagnostic accuracy as the technology does not have classic human limitations such as fatigue and difficulty concentrating [23,24]. Our experts agreed that the use of AI technology can improve diagnostic accuracy but stated that the main reason for this improvement is the fact that AI systems are better at dealing with large and complex data than humans. In addition, although some HCWs expressed hope that relying on AI systems might provide legal protections [23], our experts explained that AI-CDSSs currently cannot operate without human oversight, are sometimes inaccurate, and lack both explainability and accountability. Consequently, liability fully remains with the HCWs operating the system, a current limitation of the technology widely discussed in the literature [30,54,55].

From the interviews, several critical implementation criteria and requirements emerged. In accordance with the literature [10,30], the interviewed experts agreed that high performance is the essential criterion for implementing AI-enabled systems in health care. Ideally, deployed algorithms should outperform human experts, explain their predictions, be approved by regulatory bodies, and be frequently revalidated. Easy and unintrusive integration into existing infrastructures and workflows, intuitive and user-friendly design, and high user acceptance were frequently mentioned as essential requirements. Specifically, lack of trust and user acceptance have also been widely discussed in the existing literature as major problems

for the successful adoption of AI technology in health care [21,29]. There was consensus among our interviewees that the involvement of health care facility leaders, regulatory bodies, and end users is critical to AI adoption. Moreover, experts emphasized that users require training to interact safely with the technology. By already integrating the topic of AI in health care into the medical curriculum, users can develop the knowledge and understanding, especially of the limitations, and the confidence needed to use AI in a clinical setting [56].

The interviewees identified multiple challenges in implementing AI systems in health care. Many mentioned strict data protection and security regulations, complex certification processes, and the unresolved question of liability as fundamental regulatory challenges to technological development and deployment. These regulatory aspects have been discussed in previous research, especially from the side of policy makers. In addition, previous work has focused on ethical considerations such as the lack of transparency and discrimination in the context of AI-CDSSs in health care [21,29]. The experts agreed on several significant logistical challenges such as procuring funding for AI systems, the lack of capable IT professionals needed for technology implementation and maintenance, and difficulties with sharing and integrating data across institutions. From a technical standpoint, the lack of available preprocessed, representative, high-quality data impairs the training of high-performance AI algorithms for the entire patient population. Researchers surveyed in previous studies have confirmed these challenges and also stated that useful data are expensive and often come with computational limitations [30,31]. Our interviewees mentioned that institutions' outdated and inflexible IT infrastructures are also a big challenge for deploying AI technology. Correspondingly, IT experts in previous studies have emphasized the compatibility problems of the systems with the existing IT infrastructure [32].

The interviewed experts mentioned that implementing AI technology holds both threats and opportunities for the future. Concerns were expressed that biased training data might exacerbate health care disparities, hurting marginalized groups, and that automation bias might lead to medical errors. Moreover, AI systems in health care could become a target for cyberattacks. Previous research has shown that HCWs are concerned about losing skills and potentially even their jobs owing to AI technology. HCWs also worry about the adverse effects of using AI systems for the physician-patient relationship and patient privacy [13,23,25,26]. Our experts also acknowledged the problem of losing training-intensive skills but disagreed with the notion that AI systems will make some HCWs obsolete in the foreseeable future. In addition, they did not mention HCW-patient relationships or patient privacy as major limitations of AI systems. Overall, our experts agreed consistently with the previously mentioned opportunities that the technology could offer [8-14,23,24]: workload reduction for HCWs, improvements in diagnostic accuracy and patients' health outcomes, and advances in personalized medicine and optimized health care logistics.

Generally, the statements of both professional groups closely coincided; however, we also found some interesting differences. IT professionals emphasized China's leading role in AI

technology more strongly than researchers. In particular, the researchers blamed the state of digitization and numerous regulations for why Europe is lagging behind. Researchers emphasized the need for high security of the system and its regular validation. Some researchers recommended simply letting AI technology work in the background. However, if not, the system should integrate smoothly into the existing workflow. Consequently, researchers called for users to know how AI systems work to understand their limitations. The 2 groups also highlighted different implementation challenges—for example, IT experts considered biased training data as one of the biggest challenges. Researchers naturally focused much more on technical challenges such as data availability, technical infrastructure, and interfaces. Interestingly, only researchers mentioned overreliance on the system as a real threat from AI technology. Finally, considering future opportunities, IT experts highlighted themes such as increasing health care service availability and improving clinical outcomes, whereas researchers focused more on reducing HCWs' workloads.

After proportionally adjusting for the imbalance between respondents from Western Europe and North America, we found that their views differed on some topics. North American experts spoke more frequently and in more detail about the overarching themes of AI, machine learning, algorithms, and technology. Many European respondents felt that the lack of available and shareable data is the reason that AI development and adoption in Europe are slow. They exclusively indicated that lack of accountability and open liability issues were major limiting factors for using AI systems. Accordingly, answering these questions was a necessary criterion for implementing the technology. Interviewees from North America emphasized regular system validation and seamless workflow integration, ideally working only in the background, as critical implementation criteria. They also saw biased training data as one of the biggest threats to AI integration. Overall, North American respondents were more likely to talk about implementation challenges.

Implications for Research and Practice

In total, 5 aspects emerged from the interviews that seem to be particularly important in the context of AI implementation in health care. First, data protection is a central element for AI development and adoption as it regulates access to training data and has implications for the performance of AI support tools. The problem of a lack of available and shareable data is specifically prominent in European countries. If Europe wants to keep up with the global players in AI-enabled technology for health care, a fundamental change in the rules on how data are made available, shared, and integrated across institutions will be needed. Second, all stakeholders seemed to agree that high performance is the most fundamental aspect of successfully implementing AI systems in health care. To ensure high performance in the real world, AI systems have to be continuously monitored and revalidated in the environment in which they operate. Third, as the end users of many AI systems for health care, HCWs play an important role in the successful implementation of the technology and should be prepared accordingly. HCWs should be trained on how to interact effectively and safely with the technology and learn about its

limitations to avoid relying on incorrect advice. Research should be conducted to identify the most appropriate and effective strategy to train HCWs on the technology. Fourth, it is also striking that ethical concerns are hardly addressed except for data protection and possible biases within the data. Further development of AI systems in health care should necessarily take place within a defined ethical framework for action as the technologies are in direct contact with sensitive patient data and humans themselves. Finally, given that researchers and IT professionals often raise different issues on similar topics, it is important to ensure that all stakeholders involved in AI implementation collaborate and consider each other's opinions. AI systems should be developed to meet the needs of and support practitioners in their everyday work; consequently, their views should matter the most.

Limitations

This study has several limitations. First, the regional backgrounds of our interview partners were not perfectly balanced. Overall, more participants worked in Western Europe than in North America, with a much larger proportion of the IT experts interviewed coming from Western Europe. This might have skewed the results toward a more European-centered view. Even the proportional adjustment of the statements of the underrepresented group of experts cannot guarantee a balanced picture. Second, experts from several but not all Western European countries, let alone all European countries, were interviewed. In particular, experts from Baltic and Scandinavian countries would be of interest to the study as these regions were frequently mentioned by the interviewees as European pioneers in AI technology. In addition, the North American expert group consisted only of people who worked in the United States or Canada. Third, the focus of interviewees in the field of radiology may have been due to selection bias as several interviewees (7/23, 30%) had strong domain expertise in radiology, which is understandable as this is the field where AI technologies are

commonly used. However, some aspects relevant to implementing AI-enabled systems in other medical fields may have been overlooked. Finally, inherent features of qualitative expert interview studies (including small and, to a degree, self-selected samples and nonstandardized data analysis) cannot ensure the generalizability of the results. Subsequent studies should provide a more balanced and broader field of experts and use more quantitative methods to improve generalizability. To gain an even more global view of the current state of AI systems in health care, experts from other countries, especially China and wider parts of Europe and North America, should be included in future research.

Conclusions

Our study provides new insights into the implementation process of AI technology in health care from the perspective of AI researchers and IT professionals in North America and Western Europe. Our cross-professional and international approach revealed nuanced views on various topics from 2 stakeholder groups actively involved in the technology's deployment. Although interviewees from both groups and regions had relatively consistent views, they often focused on different aspects that they deemed most relevant. This highlights the importance of systematically documenting technology adoption expectations and challenges from different perspectives to avoid overlooking some critical elements. Our findings provide a broad overview of the current state, criteria, challenges, and prospects for the deployment of AI technology in health care. To advance the technology and make it widely available, critical implementation criteria have to be met, and all stakeholders must collaborate to overcome the challenges hindering the technology from reaching its full potential. By designing the development processes based on participatory design principles, AI-enabled applications can truly help solve current and future problems faced by health care systems worldwide.

Acknowledgments

This work was conducted with financial support from the Volkswagen Foundation (grant 98 525). The Volkswagen Foundation played no role in the study design, report writing, or decision to submit the manuscript for publication. The authors thank Eesha Kokje for copyediting the manuscript.

Conflicts of Interest

None declared.

References

1. World Health Organization. Regional Office for Europe. Health systems respond to noncommunicable diseases: time for ambition: summary. World Health Organization. 2019. URL: <https://tinyurl.com/ytz834xj> [accessed 2023-09-30]
2. World Health Organization. World health statistics 2022: monitoring health for the SDGs, sustainable development goals. World Health Organization. 2022 May 19. URL: <https://www.who.int/publications/i/item/9789240051157> [accessed 2023-09-30]
3. World Health Organization. Global spending on health: rising to the pandemic's challenges. World Health Organization. 2022 Dec 8. URL: <https://www.who.int/publications/i/item/9789240064911> [accessed 2023-09-30]
4. Arndt BG, Beasley JW, Watkinson MD, Temte JL, Tuan WJ, Sinsky CA, et al. Tethered to the EHR: primary care physician workload assessment using EHR event log data and time-motion observations. *Ann Fam Med* 2017 Sep;15(5):419-426 [FREE Full text] [doi: [10.1370/afm.2121](https://doi.org/10.1370/afm.2121)] [Medline: [28893811](https://pubmed.ncbi.nlm.nih.gov/28893811/)]

5. Pruszyński J, Cianciara D, Pruszyńska I, Włodarczyk-Pruszyńska I. Staff shortages and inappropriate work conditions as a challenge geriatrics and contemporary healthcare service at large faces. *J Educ Health Sport* 2022 May 24;12(7):136-147. [doi: [10.12775/JEHS.2022.12.07.014](https://doi.org/10.12775/JEHS.2022.12.07.014)]
6. Brborović O, Brborović H, Hrain L. The COVID-19 pandemic crisis and patient safety culture: a mixed-method study. *Int J Environ Res Public Health* 2022 Feb 16;19(4):2237 [FREE Full text] [doi: [10.3390/ijerph19042237](https://doi.org/10.3390/ijerph19042237)] [Medline: [35206429](https://pubmed.ncbi.nlm.nih.gov/35206429/)]
7. Garcia CL, Abreu LC, Ramos JL, Castro CF, Smiderle FR, Santos JA, et al. Influence of burnout on patient safety: systematic review and meta-analysis. *Medicina (Kaunas)* 2019 Aug 30;55(9):553 [FREE Full text] [doi: [10.3390/medicina55090553](https://doi.org/10.3390/medicina55090553)] [Medline: [31480365](https://pubmed.ncbi.nlm.nih.gov/31480365/)]
8. Sutton RT, Pincock D, Baumgart DC, Sadowski DC, Fedorak RN, Kroeker KI. An overview of clinical decision support systems: benefits, risks, and strategies for success. *NPJ Digit Med* 2020;3:17 [FREE Full text] [doi: [10.1038/s41746-020-0221-y](https://doi.org/10.1038/s41746-020-0221-y)] [Medline: [32047862](https://pubmed.ncbi.nlm.nih.gov/32047862/)]
9. Lee H, Yune S, Mansouri M, Kim M, Tajmir SH, Guerrier CE, et al. An explainable deep-learning algorithm for the detection of acute intracranial haemorrhage from small datasets. *Nat Biomed Eng* 2019 Mar 17;3(3):173-182. [doi: [10.1038/s41551-018-0324-9](https://doi.org/10.1038/s41551-018-0324-9)] [Medline: [30948806](https://pubmed.ncbi.nlm.nih.gov/30948806/)]
10. Topol EJ. High-performance medicine: the convergence of human and artificial intelligence. *Nat Med* 2019 Jan 7;25(1):44-56. [doi: [10.1038/s41591-018-0300-7](https://doi.org/10.1038/s41591-018-0300-7)] [Medline: [30617339](https://pubmed.ncbi.nlm.nih.gov/30617339/)]
11. Aung YY, Wong DC, Ting DS. The promise of artificial intelligence: a review of the opportunities and challenges of artificial intelligence in healthcare. *Br Med Bull* 2021 Sep 10;139(1):4-15. [doi: [10.1093/bmb/ldab016](https://doi.org/10.1093/bmb/ldab016)] [Medline: [34405854](https://pubmed.ncbi.nlm.nih.gov/34405854/)]
12. Gillan C, Milne E, Harnett N, Purdie TG, Jaffray DA, Hodges B. Professional implications of introducing artificial intelligence in healthcare: an evaluation using radiation medicine as a testing ground. *J Radiother Pract* 2018 Oct 03;18(1):5-9. [doi: [10.1017/S1460396918000468](https://doi.org/10.1017/S1460396918000468)]
13. Chen Y, Stavropoulou C, Narasinkan R, Baker A, Scarbrough H. Professionals' responses to the introduction of AI innovations in radiology and their implications for future adoption: a qualitative study. *BMC Health Serv Res* 2021 Aug 14;21(1):813 [FREE Full text] [doi: [10.1186/s12913-021-06861-y](https://doi.org/10.1186/s12913-021-06861-y)] [Medline: [34389014](https://pubmed.ncbi.nlm.nih.gov/34389014/)]
14. Blease C, Kaptchuk TJ, Bernstein MH, Mandl KD, Halamka JD, DesRoches CM. Artificial intelligence and the future of primary care: exploratory qualitative study of UK general practitioners' views. *J Med Internet Res* 2019 Mar 20;21(3):e12802 [FREE Full text] [doi: [10.2196/12802](https://doi.org/10.2196/12802)] [Medline: [30892270](https://pubmed.ncbi.nlm.nih.gov/30892270/)]
15. Thomas LB, Mastorides SM, Viswanadhan NA, Jakey CE, Borkowski AA. Artificial intelligence: review of current and future applications in medicine. *Fed Pract* 2021 Nov;38(11):527-538 [FREE Full text] [doi: [10.12788/fp.0174](https://doi.org/10.12788/fp.0174)] [Medline: [35136337](https://pubmed.ncbi.nlm.nih.gov/35136337/)]
16. Briganti G, Le Moine O. Artificial intelligence in medicine: today and tomorrow. *Front Med (Lausanne)* 2020 Feb 5;7:27 [FREE Full text] [doi: [10.3389/fmed.2020.00027](https://doi.org/10.3389/fmed.2020.00027)] [Medline: [32118012](https://pubmed.ncbi.nlm.nih.gov/32118012/)]
17. Busnatu S, Niculescu AG, Bolocan A, Petrescu GE, Păduraru DN, Năstasă I, et al. Clinical applications of artificial intelligence-an updated overview. *J Clin Med* 2022 Apr 18;11(8):2265 [FREE Full text] [doi: [10.3390/jcm11082265](https://doi.org/10.3390/jcm11082265)] [Medline: [35456357](https://pubmed.ncbi.nlm.nih.gov/35456357/)]
18. Yin J, Ngiam KY, Teo HH. Role of artificial intelligence applications in real-life clinical practice: systematic review. *J Med Internet Res* 2021 Apr 22;23(4):e25759 [FREE Full text] [doi: [10.2196/25759](https://doi.org/10.2196/25759)] [Medline: [33885365](https://pubmed.ncbi.nlm.nih.gov/33885365/)]
19. Bajgain B, Lorenzetti D, Lee J, Sauro K. Determinants of implementing artificial intelligence-based clinical decision support tools in healthcare: a scoping review protocol. *BMJ Open* 2023 Feb 23;13(2):e068373 [FREE Full text] [doi: [10.1136/bmjopen-2022-068373](https://doi.org/10.1136/bmjopen-2022-068373)] [Medline: [36822813](https://pubmed.ncbi.nlm.nih.gov/36822813/)]
20. Scott IA, Carter SM, Coiera E. Exploring stakeholder attitudes towards AI in clinical practice. *BMJ Health Care Inform* 2021 Dec 09;28(1):e100450. [doi: [10.1136/bmjhci-2021-100450](https://doi.org/10.1136/bmjhci-2021-100450)] [Medline: [34887331](https://pubmed.ncbi.nlm.nih.gov/34887331/)]
21. Asan O, Bayrak AE, Choudhury A. Artificial intelligence and human trust in healthcare: focus on clinicians. *J Med Internet Res* 2020 Jun 19;22(6):e15154 [FREE Full text] [doi: [10.2196/15154](https://doi.org/10.2196/15154)] [Medline: [32558657](https://pubmed.ncbi.nlm.nih.gov/32558657/)]
22. Khairat S, Marc D, Crosby W, Al Sanousi A. Reasons for physicians not adopting clinical decision support systems: critical analysis. *JMIR Med Inform* 2018 Apr 18;6(2):e24 [FREE Full text] [doi: [10.2196/medinform.8912](https://doi.org/10.2196/medinform.8912)] [Medline: [29669706](https://pubmed.ncbi.nlm.nih.gov/29669706/)]
23. Buck C, Doctor E, Hennrich J, Jöhnk J, Eymann T. General practitioners' attitudes toward artificial intelligence-enabled systems: interview study. *J Med Internet Res* 2022 Jan 27;24(1):e28916 [FREE Full text] [doi: [10.2196/28916](https://doi.org/10.2196/28916)] [Medline: [35084342](https://pubmed.ncbi.nlm.nih.gov/35084342/)]
24. van der Zander QE, van der Ende-van Loon MC, Janssen JM, Winkens B, van der Sommen F, Masclee AA, et al. Artificial intelligence in (gastrointestinal) healthcare: patients' and physicians' perspectives. *Sci Rep* 2022 Oct 06;12(1):16779 [FREE Full text] [doi: [10.1038/s41598-022-20958-2](https://doi.org/10.1038/s41598-022-20958-2)] [Medline: [36202957](https://pubmed.ncbi.nlm.nih.gov/36202957/)]
25. Botwe BO, Antwi WK, Arkoh S, Akudjedu TN. Radiographers' perspectives on the emerging integration of artificial intelligence into diagnostic imaging: the Ghana study. *J Med Radiat Sci* 2021 Sep 14;68(3):260-268 [FREE Full text] [doi: [10.1002/jmrs.460](https://doi.org/10.1002/jmrs.460)] [Medline: [33586361](https://pubmed.ncbi.nlm.nih.gov/33586361/)]
26. Gong B, Nugent JP, Guest W, Parker W, Chang PJ, Khosa F, et al. Influence of artificial intelligence on Canadian medical students' preference for radiology specialty: a national survey study. *Acad Radiol* 2019 Apr;26(4):566-577. [doi: [10.1016/j.acra.2018.10.007](https://doi.org/10.1016/j.acra.2018.10.007)] [Medline: [30424998](https://pubmed.ncbi.nlm.nih.gov/30424998/)]

27. Leimanis A, Palkova K. Ethical guidelines for artificial intelligence in healthcare from the sustainable development perspective. *Eur J Sustain Dev* 2021 Feb 01;10(1):90. [doi: [10.14207/ejsd.2021.v10n1p90](https://doi.org/10.14207/ejsd.2021.v10n1p90)]
28. Lennon MR, Bouamrane MM, Devlin AM, O'Connor S, O'Donnell C, Chetty U, et al. Readiness for delivering digital health at scale: lessons from a longitudinal qualitative evaluation of a national digital health innovation program in the United Kingdom. *J Med Internet Res* 2017 Feb 16;19(2):e42 [FREE Full text] [doi: [10.2196/jmir.6900](https://doi.org/10.2196/jmir.6900)] [Medline: [28209558](https://pubmed.ncbi.nlm.nih.gov/28209558/)]
29. Cutillo CM, Sharma KR, Foschini L, Kundu S, Mackintosh M, Mandl KD, et al. Machine intelligence in healthcare—perspectives on trustworthiness, explainability, usability, and transparency. *NPJ Digit Med* 2020 Mar 26;3(1):47 [FREE Full text] [doi: [10.1038/s41746-020-0254-2](https://doi.org/10.1038/s41746-020-0254-2)] [Medline: [32258429](https://pubmed.ncbi.nlm.nih.gov/32258429/)]
30. Rajpurkar P, Chen E, Banerjee O, Topol EJ. AI in health and medicine. *Nat Med* 2022 Jan 20;28(1):31-38. [doi: [10.1038/s41591-021-01614-0](https://doi.org/10.1038/s41591-021-01614-0)] [Medline: [35058619](https://pubmed.ncbi.nlm.nih.gov/35058619/)]
31. Ghassemi M, Naumann T, Schulam P, Beam AL, Chen IY, Ranganath R. A review of challenges and opportunities in machine learning for health. *AMIA Jt Summits Transl Sci Proc* 2020;2020:191-200 [FREE Full text] [Medline: [32477638](https://pubmed.ncbi.nlm.nih.gov/32477638/)]
32. Weinert L, Müller J, Svensson L, Heinze O. Perspective of information technology decision makers on factors influencing adoption and implementation of artificial intelligence technologies in 40 German hospitals: descriptive analysis. *JMIR Med Inform* 2022 Jun 15;10(6):e34678 [FREE Full text] [doi: [10.2196/34678](https://doi.org/10.2196/34678)] [Medline: [35704378](https://pubmed.ncbi.nlm.nih.gov/35704378/)]
33. Weinert L, Klass M, Schneider G, Heinze O. Exploring stakeholder requirements to enable research and development of artificial intelligence algorithms in a hospital-based generic infrastructure: results of a multistep mixed methods study. *JMIR Form Res* 2023 Apr 18;7:e43958 [FREE Full text] [doi: [10.2196/43958](https://doi.org/10.2196/43958)] [Medline: [37071450](https://pubmed.ncbi.nlm.nih.gov/37071450/)]
34. Mou X. Artificial intelligence: investment trends and selected industry uses. The World Bank. 2019 Nov 06. URL: <https://tinyurl.com/2wzr75s4> [accessed 2023-09-30]
35. Secinaro S, Calandra D, Secinaro A, Muthurangu V, Biancone P. The role of artificial intelligence in healthcare: a structured literature review. *BMC Med Inform Decis Mak* 2021 Apr 10;21(1):125 [FREE Full text] [doi: [10.1186/s12911-021-01488-9](https://doi.org/10.1186/s12911-021-01488-9)] [Medline: [33836752](https://pubmed.ncbi.nlm.nih.gov/33836752/)]
36. Dumbach P, Liu R, Jalowski M, Eskofier BM. The adoption of artificial intelligence in SMEs - a cross-national comparison in German and Chinese healthcare. In: *Proceedings of the 20th International Conference on Perspectives in Business Informatics Research (BIR 2021) Workshops (ILOG 2021)*. 2021 Presented at: 20th International Conference on Perspectives in Business Informatics Research (BIR 2021) Workshops (ILOG 2021); September 22-24, 2021; Vienna, Austria URL: <https://ceur-ws.org/Vol-2991/paper08.pdf>
37. Owen R, Pansera M. Responsible innovation and responsible research and innovation. In: *Handbook on Science and Public Policy*. Cheltenham, UK: Edward Elgar Publishing; Jun 28, 2019.
38. Marshall MN. Sampling for qualitative research. *Fam Pract* 1996 Dec;13(6):522-525. [doi: [10.1093/fampra/13.6.522](https://doi.org/10.1093/fampra/13.6.522)] [Medline: [9023528](https://pubmed.ncbi.nlm.nih.gov/9023528/)]
39. Fereday J, Muir-Cochrane E. Demonstrating rigor using thematic analysis: a hybrid approach of inductive and deductive coding and theme development. *Int J Qual Methods* 2016 Nov 29;5(1):80-92. [doi: [10.1177/160940690600500107](https://doi.org/10.1177/160940690600500107)]
40. Crabtree BF, Miller, WF. A template approach to text analysis: developing and using codebooks. In: *Doing Qualitative Research*. Thousand Oaks, CA: SAGE Publications; 1992.
41. Boyatzis RE. *Transforming Qualitative Information Thematic Analysis and Code Development*. Thousand Oaks, CA: SAGE Publications; Apr 16, 1998.
42. Hummelsberger P, Koch T, Rauh S, Dorn J, Lermer E, Raue M, et al. Insights on the current state and future outlook of artificial intelligence in healthcare from expert interviews. OSF Home Preprint posted online July 5, 2023. [FREE Full text] [doi: [10.17605/OSF.IO/GEU26](https://doi.org/10.17605/OSF.IO/GEU26)]
43. Bluhm DJ, Harman W, Lee TW, Mitchell TR. Qualitative research in management: a decade of progress. *J Manag Stud* 2011 Oct 18;48(8):1866-1891. [doi: [10.1111/j.1467-6486.2010.00972.x](https://doi.org/10.1111/j.1467-6486.2010.00972.x)]
44. Wilhelmy A, Kleinmann M, König CJ, Melchers KG, Truxillo DM. How and why do interviewers try to make impressions on applicants? a qualitative study. *J Appl Psychol* 2016 Mar;101(3):313-332. [doi: [10.1037/apl0000046](https://doi.org/10.1037/apl0000046)] [Medline: [26436440](https://pubmed.ncbi.nlm.nih.gov/26436440/)]
45. Banerjee M, Capozzoli M, McSweeney L, Sinha D. Beyond kappa: a review of interrater agreement measures. *Can J Stat* 1999 Mar;27(1):3-23. [doi: [10.2307/3315487](https://doi.org/10.2307/3315487)]
46. Landis JR, Koch GG. The measurement of observer agreement for categorical data. *Biometrics* 1977 Mar;33(1):159-174. [doi: [10.2307/2529310](https://doi.org/10.2307/2529310)]
47. Jacobs T, Tschötschel R. Topic models meet discourse analysis: a quantitative tool for a qualitative approach. *Int J Soc Res Methodol* 2019 Feb 07;22(5):469-485. [doi: [10.1080/13645579.2019.1576317](https://doi.org/10.1080/13645579.2019.1576317)]
48. Leeson W, Resnick A, Alexander D, Rovers J. Natural language processing (NLP) in qualitative public health research: a proof of concept study. *Int J Qual Methods* 2019 Nov 13;18:160940691988702. [doi: [10.1177/1609406919887021](https://doi.org/10.1177/1609406919887021)]
49. Miyaoka A, Decker-Woodrow L, Hartman N, Booker B, Ottmar E. Emergent coding and topic modeling: a comparison of two qualitative analysis methods on teacher focus group data. *Int J Qual Methods* 2023 Mar 22;22:160940692311659. [doi: [10.1177/16094069231165950](https://doi.org/10.1177/16094069231165950)]
50. Campbell JC, Hindle A, Stroulia E. Latent dirichlet allocation: extracting topics from software engineering data. In: *The Art and Science of Analyzing Software Data*. Burlington, MA: Morgan Kaufmann; 2016:139-159.

51. Witten IH, Frank E, Hall MA, Pal CJ. Data Mining: Practical Machine Learning Tools and Techniques. Amsterdam, The Netherlands: Elsevier Science; Oct 2016.
52. Straka M, Straková J. Tokenizing, POS Tagging, Lemmatizing and Parsing UD 2.0 with UDPipe. In: Proceedings of the CoNLL 2017 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies. 2017 Presented at: CoNLL 2017 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies; August 3-4, 2017; Vancouver, BC. [doi: [10.18653/v1/k17-3009](https://doi.org/10.18653/v1/k17-3009)]
53. Grün B, Hornik K. topicmodels: an R package for fitting topic models. J Stat Softw 2011;40(13):1-30. [doi: [10.18637/jss.v040.i13](https://doi.org/10.18637/jss.v040.i13)]
54. Jassar S, Adams SJ, Zarzeczny A, Burbridge BE. The future of artificial intelligence in medicine: medical-legal considerations for health leaders. Health Manage Forum 2022 May 31;35(3):185-189 [FREE Full text] [doi: [10.1177/08404704221082069](https://doi.org/10.1177/08404704221082069)] [Medline: [35354409](https://pubmed.ncbi.nlm.nih.gov/35354409/)]
55. Rowland SP, Fitzgerald JE, Lungren M, Lee EH, Harned Z, McGregor AH. Digital health technology-specific risks for medical malpractice liability. NPJ Digit Med 2022 Oct 20;5(1):157 [FREE Full text] [doi: [10.1038/s41746-022-00698-3](https://doi.org/10.1038/s41746-022-00698-3)] [Medline: [36261469](https://pubmed.ncbi.nlm.nih.gov/36261469/)]
56. Weidener L, Fischer M. Artificial intelligence teaching as part of medical education: qualitative analysis of expert interviews. JMIR Med Educ 2023 Apr 24;9:e46428 [FREE Full text] [doi: [10.2196/46428](https://doi.org/10.2196/46428)] [Medline: [36946094](https://pubmed.ncbi.nlm.nih.gov/36946094/)]

Abbreviations

AI: artificial intelligence

AI-CDSS: artificial intelligence-enabled clinical decision support system

EHR: electronic health record

FDA: Food and Drug Administration

GDPR: General Data Protection Regulation

HCW: health care worker

ITEU: IT expert in Western Europe

ITNA: IT expert in North America

REEU: researcher in Western Europe

RENA: researcher in North America

Edited by K El Emam, B Malin; submitted 16.03.23; peer-reviewed by C Wang, W LaMendola, H Heppner, L Weinert; comments to author 14.05.23; revised version received 06.07.23; accepted 01.08.23; published 31.10.23.

Please cite as:

Hummelsberger P, Koch TK, Rauh S, Dorn J, Lerner E, Raue M, Hudecek MFC, Schicho A, Colak E, Ghassemi M, Gaube S

Insights on the Current State and Future Outlook of AI in Health Care: Expert Interview Study

JMIR AI 2023;2:e47353

URL: <https://ai.jmir.org/2023/1/e47353>

doi: [10.2196/47353](https://doi.org/10.2196/47353)

PMID: [38875571](https://pubmed.ncbi.nlm.nih.gov/38875571/)

©Pia Hummelsberger, Timo K Koch, Sabrina Rauh, Julia Dorn, Eva Lerner, Martina Raue, Matthias F C Hudecek, Andreas Schicho, Errol Colak, Marzyeh Ghassemi, Susanne Gaube. Originally published in JMIR AI (<https://ai.jmir.org>), 31.10.2023. This is an open-access article distributed under the terms of the Creative Commons Attribution License (<https://creativecommons.org/licenses/by/4.0/>), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR AI, is properly cited. The complete bibliographic information, a link to the original publication on <https://www.ai.jmir.org/>, as well as this copyright and license information must be included.

Original Paper

Machine Learning Models Versus the National Early Warning Score System for Predicting Deterioration: Retrospective Cohort Study in the United Arab Emirates

Hazem Lashen¹, BSc; Terrence Lee St John², PhD; Y Zaki Almallah², MD; Madhu Sasidhar³, MD; Farah E Shamout¹, DPhil

¹Engineering Division, New York University Abu Dhabi, Abu Dhabi, United Arab Emirates

²Cleveland Clinic Abu Dhabi, Abu Dhabi, United Arab Emirates

³Cleveland Clinic Tradition Hospital, Port St. Lucie, FL, United States

Corresponding Author:

Farah E Shamout, DPhil

Engineering Division

New York University Abu Dhabi

Saadiyat Island

Abu Dhabi, 129188

United Arab Emirates

Phone: 971 2 6284184

Email: farah.shamout@nyu.edu

Abstract

Background: Early warning score systems are widely used for identifying patients who are at the highest risk of deterioration to assist clinical decision-making. This could facilitate early intervention and consequently improve patient outcomes; for example, the National Early Warning Score (NEWS) system, which is recommended by the Royal College of Physicians in the United Kingdom, uses predefined alerting thresholds to assign scores to patients based on their vital signs. However, there is limited evidence of the reliability of such scores across patient cohorts in the United Arab Emirates.

Objective: Our aim in this study was to propose a data-driven model that accurately predicts in-hospital deterioration in an inpatient cohort in the United Arab Emirates.

Methods: We conducted a retrospective cohort study using a real-world data set that consisted of 16,901 unique patients associated with 26,073 inpatient emergency encounters and 951,591 observation sets collected between April 2015 and August 2021 at a large multispecialty hospital in Abu Dhabi, United Arab Emirates. The observation sets included routine measurements of heart rate, respiratory rate, systolic blood pressure, level of consciousness, temperature, and oxygen saturation, as well as whether the patient was receiving supplementary oxygen. We divided the data set of 16,901 unique patients into training, validation, and test sets consisting of 11,830 (70%; 18,319/26,073, 70.26% emergency encounters), 3397 (20.1%; 5206/26,073, 19.97% emergency encounters), and 1674 (9.9%; 2548/26,073, 9.77% emergency encounters) patients, respectively. We defined an adverse event as the occurrence of admission to the intensive care unit, mortality, or both if the patient was admitted to the intensive care unit first. On the basis of 7 routine vital signs measurements, we assessed the performance of the NEWS system in detecting deterioration within 24 hours using the area under the receiver operating characteristic curve (AUROC). We also developed and evaluated several machine learning models, including logistic regression, a gradient-boosting model, and a feed-forward neural network.

Results: In a holdout test set of 2548 encounters with 95,755 observation sets, the NEWS system achieved an overall AUROC value of 0.682 (95% CI 0.673-0.690). In comparison, the best-performing machine learning models, which were the gradient-boosting model and the neural network, achieved AUROC values of 0.778 (95% CI 0.770-0.785) and 0.756 (95% CI 0.749-0.764), respectively. Our interpretability results highlight the importance of temperature and respiratory rate in predicting patient deterioration.

Conclusions: Although traditional early warning score systems are the dominant form of deterioration prediction models in clinical practice today, we strongly recommend the development and use of cohort-specific machine learning models as an alternative. This is especially important in external patient cohorts that were unseen during model development.

KEYWORDS

machine learning; early warning score system; clinical deterioration; early warning; score system; cohort; real-world data; neural network; predict; deterioration

Introduction**Background**

Early warning score (EWS) systems are a staple of modern clinical practice because they provide a standardized method for detecting in-hospital patient deterioration. Several other systems have been introduced with the advent of computerized medical records [1,2], such as the Modified Early Warning Score system [3] and the National Early Warning Score (NEWS) system [4], which is recommended by the Royal College of Physicians in the United Kingdom.

Such systems assign an overall aggregate score to the patient to indicate their overall risk of deterioration, based on a predetermined set of alerting ranges [5]; for example, the alerting thresholds of the NEWS system are shown in Table 1. Later work introduced EWS systems tailored for specific patient subgroups, such as for pediatrics [6] or cardiovascular-related deterioration [7]. The main strengths of EWS systems are that they are simple, easy to use, and highly interpretable [8,9], which facilitates their use in hospitals, including those with limited resources [10,11].

Table 1. Summary of the National Early Warning Score system, with the thresholds of the system outlined. For a given set of vital signs measurements, each variable is compared against its respective threshold and assigned a score accordingly. The patient's overall score is the summation of scores assigned to all variables.

Vital sign	Score						
	3	2	1	0	1	2	3
Heart rate (beats/min)	≤40	N/A ^a	41-50	51-90	91-110	111-130	≥131
Oxygen saturation (%)	≤91	92-93	94-95	≥96	N/A	N/A	N/A
Temperature (°C)	≤35	N/A	35.1-36.0	36.1-38.0	38.1-39.0	≥39.1	N/A
Systolic blood pressure (mm Hg)	≤90	91-100	101-110	111-219	N/A	N/A	≥220
Respiratory rate (breaths/min)	≤8	N/A	9-11	12-20	N/A	21-24	≥25
Level of consciousness	N/A	N/A	N/A	Alert	N/A	N/A	Voice, pain, or unresponsive
Supplementary oxygen	N/A	Yes	N/A	No	N/A	N/A	N/A

^aN/A: not applicable.

Despite their ubiquity, EWS systems also have limitations. Many of the alerting thresholds are defined in a heuristic manner with respect to a specific deterioration timeline, which makes it increasingly difficult to modify the thresholds for cohorts with significantly different characteristics or demographics than those relied upon during model development, as witnessed during the COVID-19 pandemic [12-14]. In addition, EWS systems do not capture any relationships between the input variables and commonly treat them equally, despite some being more indicative of deterioration than others [1]. However, because of their simplicity, they have been widely deployed in hospitals around the world.

In recent years, machine learning (ML) techniques have gained popularity in the development of deterioration prediction models [15-18] by treating the problem as a binary classification task [19-21]. Such approaches range from gradient-boosted trees [12,22], which consist of an ensemble of tree models, to neural networks (NNs) [21,23] and have been used in different scenarios where deterioration prediction is needed [24-26]. Although ML models have been shown to outperform traditional EWS systems [19,27], especially during the COVID-19

pandemic [28,29], one of their main limitations is the lack of interpretability compared with traditional EWS systems [30].

Objectives

Our aim in this study was to propose a data-driven model that predicts patient deterioration with high accuracy in an inpatient cohort in Abu Dhabi, United Arab Emirates. To this end, we assessed and compared the performance of the NEWS system with that of 3 ML models, namely logistic regression (LR), gradient-boosted trees, and NNs, and developed and evaluated the models using a real-world data set collected at a multispecialty hospital in Abu Dhabi. We also used Shapley additive explanations (SHAP) analysis as a way to interpret the predictions of the ML models.

Methods

This study is reported in accordance with the Transparent Reporting of a Multivariable Prediction Model for Individual Prognosis or Diagnosis (TRIPOD) guidelines [31]. The TRIPOD checklist can be found in [Multimedia Appendix 1](#).

Ethics Approval

The study received approval from the research ethics committees at Cleveland Clinic Abu Dhabi (A-2020-102) and New York University Abu Dhabi (HRPP-2020-55).

Data Set

We obtained a data set collected between April 2015 and August 2021 at the multispecialty facility Cleveland Clinic Abu Dhabi in Abu Dhabi, United Arab Emirates. The data set included patient demographics; vital signs measurements; and time stamps relating to admission to the intensive care unit (ICU) and mortality, which are the adverse events of interest in this study.

We defined the inclusion and exclusion criteria following the standard in previous work for the development of EWS systems [2] (Figure 1). First, we grouped a set of vital signs measurements to represent a single observation set if they had been recorded within the same patient encounter and shared the same time of measurement. We excluded any patients with missing identifiers or necessary information such as records pertaining to whether the patient was alive at the time of discharge, patient age, and patient or encounter identifiers, as

well as time stamps of vital signs measurements. We included inpatient admissions and excluded encounters of patients aged <18 years at the time of admission. We only included emergency encounters and excluded other types of admissions. Within each encounter, we dropped any vital signs measurements recorded after the occurrence of an adverse event, which is essentially admission to the ICU. We excluded any observation sets that contained ≥ 1 implausible observations or >2 missing vital signs measurements. An illustration of our data set processing pipeline is shown in Figure 2. The plausible ranges used are presented in Table S1 in Multimedia Appendix 2.

Finally, we split the data set randomly into training, validation, and test sets in a ratio of 7:2:1, respectively. This split was carried out on a patient level such that all examples belonging to a single patient were assigned to a single split only. We split the data randomly because we assumed that most of the patients admitted in 2020 and 2021 were patients with COVID-19 infection (the COVID-19 outbreak began approximately in March 2020 in the United Arab Emirates); therefore, we were interested in assessing the average performance of the models over time. We conducted a secondary analysis where we split the data based on time to understand the impact of a temporal split.

Figure 1. Application of the inclusion and exclusion criteria. We illustrate here the results of applying the inclusion and exclusion criteria, where p, e, and n represent the number of patients, encounters, and observation sets, respectively. We first excluded patients with missing information, such as age or patient identifiers. We included inpatient encounters of adult patients (aged >18 years) and excluded nonemergency encounters. Finally, we excluded observation sets recorded after an adverse event (AE) had occurred as well as observation sets with ≥ 1 implausible observations.

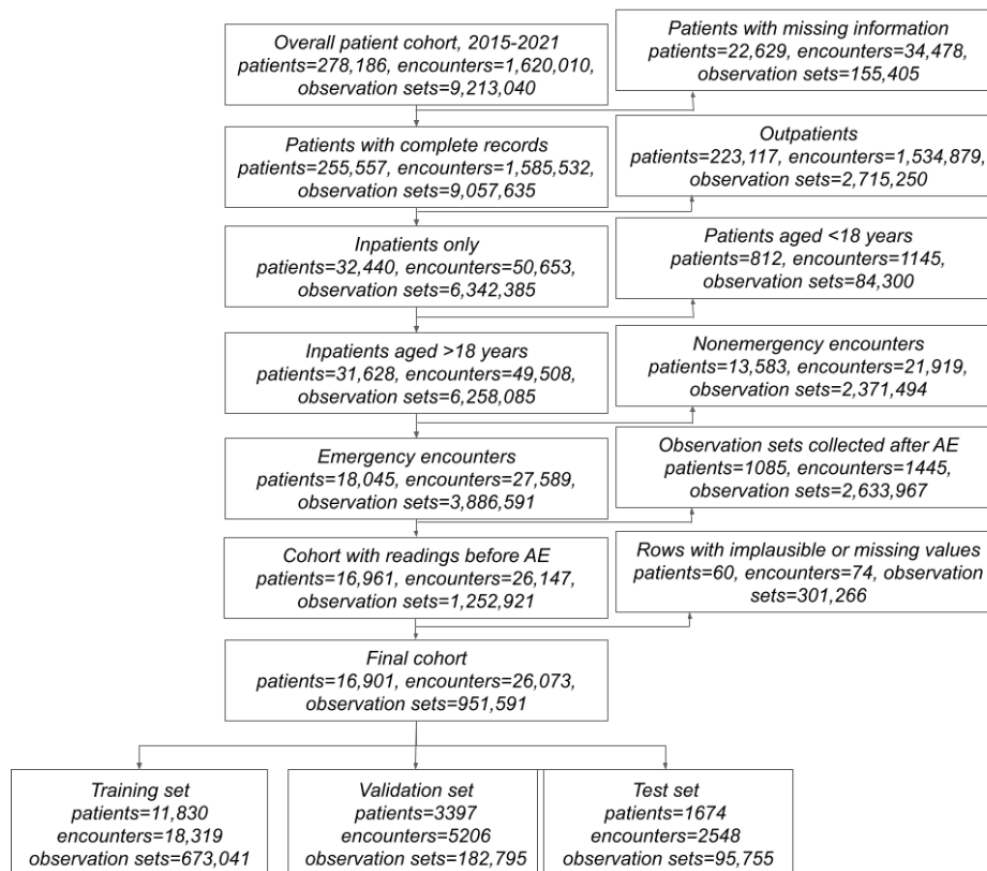
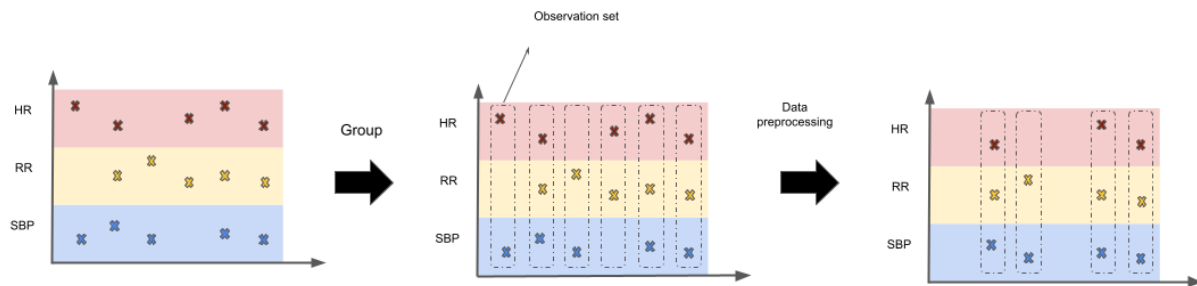


Figure 2. Definition of observation sets. We illustrate here a simplified version of how we defined observation sets using 3 vital signs only: heart rate (HR), respiratory rate (RR), and systolic blood pressure (SBP). Vital signs taken at the same time were grouped together for the same patient encounter into observation sets. We subsequently applied our inclusion and exclusion criteria to select the relevant patient encounters and associated observation sets.



Input Features

We extracted 7 vital signs variables that are used in the NEWS system. These included heart rate; respiratory rate; temperature; systolic blood pressure; oxygen saturation; level of consciousness indicated by the alert, voice, pain, or unresponsive score; and whether a patient was receiving supplementary oxygen. To derive the supplementary oxygen variable, we relied on the patient's fraction of inspired oxygen reading. We assumed that any fraction of inspired oxygen measurement of $>21\%$ indicated that the patient was receiving supplementary oxygen [32]. For level of consciousness, we used a provided binary feature in the data set that follows the scoring of the NEWS system (0 if a patient is alert and 3 otherwise). We applied mean imputation to all features, except for supplementary oxygen and level of consciousness, where a missing value was treated as not receiving supplementary oxygen and alert, respectively. We treated vital signs measurements recorded at the same time as a single observation set (Figure 2), meaning that each encounter (patient stay) contained multiple observation sets recorded at various times during the patient stay.

Outcome Definition

We defined the composite outcome of admission to the ICU and mortality as a deterioration (adverse) event. In cases of multiple adverse events, we considered the time of the first occurring event. For a given observation set, we generated binary ground-truth labels based on whether an adverse event occurred within a certain time window from the measurement time of the respective observation set. If it did indeed occur within the time window, we set the label as 1 (positive label); otherwise, we set it as 0 (negative label). To evaluate the performance of the models over different time windows, we considered 4 different values: 6, 12, 24, and 36 hours. We note that 24 hours is the standard window of evaluation in the existing literature [33].

Prediction Models

We developed several prediction models (refer to the following subsections) based on prevalent ML techniques. All models, except for the NEWS system, are fitted on the training set and optimized via hyperparameter tuning on the validation set, with final results being reported on the test set.

NEWS System

The NEWS system [4] was developed by the Royal College of Physicians to provide a standardized EWS system to easily and quickly identify patients at high risk of deterioration. The NEWS system assigns a score to 7 vital signs measurements based on predetermined alerting thresholds (Table 1). The higher the final score, the greater the risk of deterioration. For each observation set, we calculated the total score based on the scores assigned to each vital sign. We then normalized each score by dividing it by the maximum possible NEWS score, which is 20, to compute performance metrics.

Gradient-Boosting Model

We developed a gradient-boosting model, extreme gradient boosting (XGBoost) [22,34], that uses an ensemble of decision trees. We implemented this model using the *XGBoost* package [34].

LR Model

LR [35] is a simple statistical method that assumes a linear combination of the input variables and uses a sigmoid activation to compute predictions in the range between 0 and 1. We implemented this model using the *scikit learn* package [36].

NN Model

We implemented a feed-forward NN [37] consisting of 10 linear layers with scaled exponential linear unit activation function [38], followed by batch normalization to reduce overfitting. The outputs of the final layer are fed to a sigmoid activation function, which outputs predictions in the range between 0 and 1. For this model, we applied min-max normalization to the input features first, whereby the minimum and maximum values were defined using the training set for all data splits. We implemented this baseline using the PyTorch framework [39].

Evaluation Metrics

We evaluated all models using 2 main evaluation metrics: the area under the receiver operating characteristic curve (AUROC) and the area under the precision-recall curve (AUPRC). Both metrics are represented as a single number between 0 and 1 to summarize the performance of a binary classifier. The receiver operating characteristic curve plots the true positive rate against the false positive rate at different classification thresholds and indicates the model's ability to discriminate between positive and negative classes. The precision-recall curve plots precision

against recall at different classification thresholds and gives an indication of the model's average precision. The baseline performance of a random classifier is equivalent to 0.5 for the AUROC and to the ratio of positive samples to the total number of samples for the AUPRC. We computed 95% CIs for the AUROC and AUPRC metrics of all models through bootstrapping with 1000 runs [40].

In addition, we compared the difference in performance, in terms of the AUROC and AUPRC values, between each ML model and the NEWS system. We report the difference and its 95% CIs using bootstrapping with 1000 runs. We computed *P* values for each comparison using the 1-tailed permutation test with 10,000 iterations [41]. All results are reported on the test set.

Model Selection

To develop the ML models, we used random hyperparameter search [42] to select the best hyperparameters using the validation set (XGBoost, LR, and NN). We have summarized the sampling ranges of the hyperparameters in Table S2 in [Multimedia Appendix 2](#). We ran each model 10 times with hyperparameters selected randomly from predetermined ranges. We then selected the model with the hyperparameters that achieved the best performance on the validation set in terms of the AUPRC value because it is considered a more informative metric owing to class imbalance [43,44]. We trained the NN models for 250 epochs. We report the performance on the test set for the selected best models.

Model Interpretability

We used the open-source SHAP package [45] to analyze feature importance using SHAP values for the best-performing model in terms of the overall AUROC. We calculated the SHAP value for each feature such that the magnitude of the SHAP value indicates greater importance for the model's prediction, and we present the average of the absolute SHAP values for each of the 7 input features in the test set. We also present the SHAP plots for the observation sets with the highest and lowest prediction scores in the best-performing model. In addition, for each input feature, we plotted the SHAP partial dependence plot, and calculated the Pearson correlation coefficient and the Spearman rank correlation coefficient between the feature values and their respective sets of SHAP values. The partial dependence plots show the relationship between the average SHAP value and each possible vital signs measurement, whereas the coefficients indicate the overall correlation between the SHAP values and the input feature values. We also included the LR coefficients and odds ratios as a comparison point owing to the simplicity

of the LR model and the significance of the coefficients in summarizing the effect of each feature on the overall prediction of the model compared with the SHAP values.

Results

Patient Cohort

We have summarized the results of applying the inclusion and exclusion criteria in [Figure 1](#). Our data set comprised 1,620,010 encounters from 278,186 patients, yielding a total of 9,213,040 observation sets. Of the 278,186 patients, 255,557 (91.87%) had complete identifying information recorded in the data set, leading to the exclusion of the rest (22,629/278,186, 8.13%), leaving 97.87% (1,585,532/1,620,010) of the encounters and 98.31% (9,057,635/9,213,040) of the observation sets. Our study specifically targets inpatients; therefore, of the 255,557 patients, after excluding 223,117 (87.31%) outpatient encounters, 32,440 (12.69%) remained. Of these 32,440 patients, 31,628 (96.39%) were aged >18 years and thus eligible for inclusion (6,258,085/9,057,635, 69.09% observation sets recorded within 49,508/1,585,562, 3.12% encounters). Furthermore, we included only emergency encounters; thus, of the 31,268 patients, 18,045 (57.71%) were included (3,886,591/6,258,085, 62.11% observation sets recorded within 27,589/49,508, 55.73% encounters). We then excluded any observation sets that occurred after an adverse event, which meant that, of the 3,886,591 observation sets, 1,252,921 (32.24%) remained. Finally, of the 1,252,921 observation sets, we removed 301,266 (24.05%) that contained implausible readings for their respective vital signs, leaving 951,655 (75.95%) observation sets. Thus, of the 18,045 patients, 16,901 (93.66%) remained in the final cohort (associated with 26,073/27,589, 94.51% encounters recorded between April 2015 and August 2021). We divided the data set of 16,901 patients as follows: training set: 11,830 (70%; 18,319/26,073, 70.26% encounters), validation set: 3397 (20.1%; 5206/26,073, 19.97% encounters), and test set: 1674 (9.9%; 2548/26,073, 9.77% encounters).

We provide a summary of the cohort's characteristics, distributions of vital signs measurements, and occurrences of adverse events in [Table 2](#). We observed an average age of 55.3 (SD 19.3), 54.9 (SD 18.7), and 53.6 (SD 19.1) years across the training, validation, and test splits, respectively. We observed a higher proportion of male patients than female patients across all splits, with the training set comprising 59.64% (7056/11,830) male patients and the validation and testing sets comprising 60.17% (2044/3397) and 58.06% (972/1674) of male patients, respectively.

Table 2. Patient cohort summary. We provide a summary of the patient cohort characteristics across the training, validation, and test sets. This includes the patient demographics, distributions of input features, and the prevalence of the deterioration labels across the different time windows.

Characteristic	Training set	Validation set	Test set
Cohort demographics			
Patients (n=16,901), n (%)	11,830 (70)	3397 (20.1)	1674 (9.9)
Male patients ^a	7056 (59.6)	2044 (60.2)	972 (58.1)
Encounters (n=26,073), n (%)	18,319 (70.3)	5206 (20)	2548 (9.8)
Age group (years), mean (SD)	55.3 (19.3)	54.9 (18.7)	53.6 (19.1)
<40, n (%) ^b	4843 (26.4)	1371 (26.3)	672 (26.4)
40-59, n (%) ^b	5313 (29)	1506 (28.9)	728 (28.6)
≥60, n (%) ^b	8163 (44.6)	2329 (44.7)	1148 (45.1)
Encounters with composite outcome, n (%) ^b	3979 (21.7)	1144 (22)	549 (21.5)
Encounters during the COVID-19 pandemic, n (%) ^b	5594 (30.5)	1657 (31.8)	836 (32.8)
Observation sets (n=951,591), n (%)	673,041 (70.7)	182,795 (19.2)	95,755 (10.1)
Heart rate (beats/min), mean (SD; IQR)	78 (15.9; 68-90)	78 (16.3; 68-89)	79 (15.8; 70-89)
Respiratory rate (breaths/min), mean (SD; IQR)	18 (2.8; 18-20)	18 (2.9; 18-20)	18 (2.9; 18-20)
Systolic blood pressure (mm Hg), mean (SD; IQR)	122 (20.9; 109-137)	124 (21.6; 110-139)	122 (20.2; 109-136)
Temperature (°C), mean (SD; IQR)	36.7 (0.4; 36.5-36.9)	36.7 (0.5; 36.5-36.9)	36.7 (0.4; 36.5-36.9)
Oxygen saturation (%), (SD; IQR)	99 (2.0; 97-100)	99 (2.0; 97-100)	99 (2.1; 97-100)
Level of consciousness, n (%)^c			
Alert	537,853 (98.1)	144,014 (97.9)	76,376 (97.5)
Voice, pain, or unresponsive	10,685 (1.9)	3086 (2.1)	1940 (2.5)
Supplementary oxygen, n (%)^d			
Provided	15,201 (2.3)	3332 (1.8)	2086 (2.2)
Not provided	657,840 (97.7)	179,463 (98.2)	93,669 (97.8)
Deterioration, n			
Within 36 hours	36,760	10,681	5373
Death	1100	266	154
ICU ^e admission	36,255	10,571	5319
Within 24 hours	31,431	9306	4556
Death	715	127	71
ICU admission	31,088	9233	4521
Within 12 hours	25,382	7633	3658
Death	358	70	21
ICU admission	25,199	7589	3643
Within 6 hours	21,332	6476	2987
Death	166	37	9
ICU admission	21,227	6452	2979

^aTraining set: n=11,830; validation set: n=3397; test set: n=1674.

^bTraining set: n=18,319; validation set: n=5206; test set: n=2548.

^cTraining set: n=548,538; validation set: n=147,100; test set: n=78,316.

^dTraining set: n=673,041; validation set: n=182,795; test set: n=95,755.

^eICU: intensive care unit.

Performance Compared With the NEWS System

We summarize the performances of the ML models and NEWS system in Table 3 for deterioration within 24 hours in terms of AUROC and AUPRC values. We note that the NN and XGBoost models achieved the best performance. The XGBoost model achieved an AUROC value of 0.778 (95% CI 0.770-0.785) across the entire test set. The NEWS system achieved an AUROC value of 0.682 (95% CI 0.673-0.690), which means that the XGBoost model achieved an improvement of 0.096 (95% CI 0.088-0.103; $P < .001$). In terms of the AUPRC values, compared with the NEWS system, the XGBoost model achieved an improvement of 0.093 (95% CI 0.083-0.101; $P < .001$). The NN model achieved an AUROC value of 0.756 (95% CI 0.749-0.764) and an AUPRC value of 0.222 (95% CI 0.211-0.235), leading to improvements of 0.074 (95% CI 0.067-0.081; $P < .001$) and 0.061 (95% CI 0.049-0.073; $P < .001$) in AUROC and AUPRC values, respectively, compared with

the NEWS system. The LR model did not perform better than the NEWS system in terms of AUROC values, and it achieved slightly better performance in terms of AUPRC values.

In Figure 3, we show the AUROC and AUPRC results for all models on the test set when varying the lengths of the prediction time window as follows: 6, 12, 24, and 36 hours. We noted that the XGBoost model and the NN model performed best across all time windows, with a better performance by the XGBoost model in terms of both AUROC and AUPRC values across all time windows. We also noted a comparable performance between the NEWS system and the LR model, with the NEWS system achieving a superior AUROC value and the LR model achieving a better AUPRC value. In addition, the performance of all models decreased as the prediction time window increased. This likely indicates that the difficulty of the task increases as the adverse events occur further away in time.

Table 3. Model performance across different subgroups. We report performances in terms of area under the receiver operating characteristic curve (AUROC) and area under the precision-recall curve (AUPRC) values for deterioration within 24 hours in the test set. We also provide 95% CIs computed using bootstrapping.

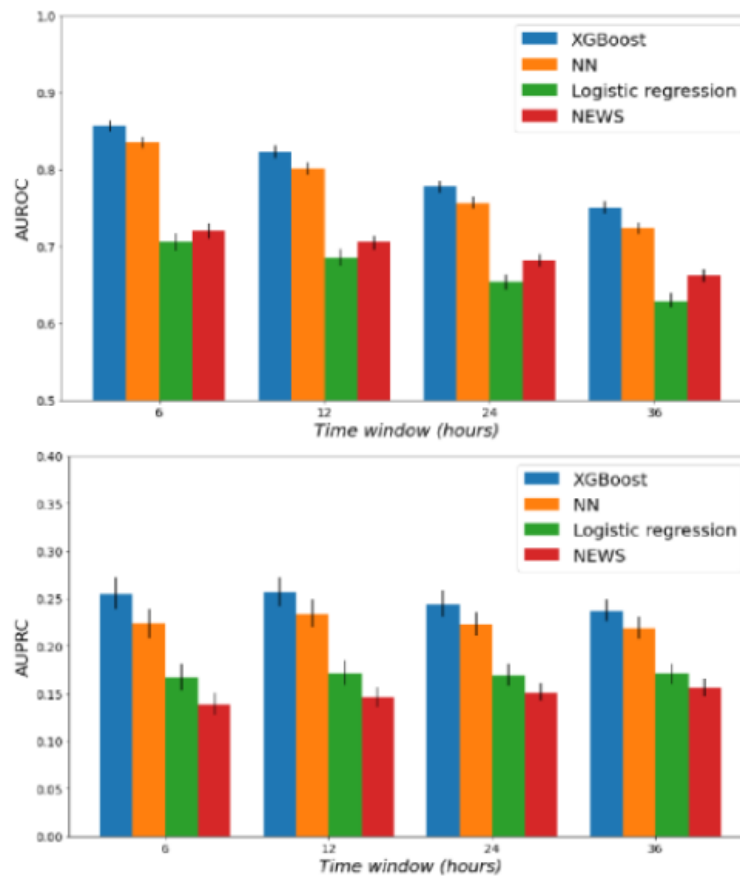
Subgroup	XGBoost ^a		Logistic regression		Neural network		NEWS ^b	
	AUROC (95% CI)	AUPRC (95% CI)	AUROC (95% CI)	AUPRC (95% CI)	AUROC (95% CI)	AUPRC (95% CI)	AUROC (95% CI)	AUPRC (95% CI)
All patients	<i>0.778^c</i> (0.770-0.785)	<i>0.244</i> (0.231-0.258)	0.654 (0.644-0.663)	0.169 (0.158-0.181)	0.756 (0.749-0.764)	0.222 (0.211-0.235)	0.682 (0.673-0.690)	0.151 (0.142-0.161)
Male patients	<i>0.775</i> (0.764-0.784)	<i>0.274</i> (0.258-0.291)	0.651 (0.638-0.663)	0.208 (0.193-0.223)	0.752 (0.742-0.762)	0.253 (0.237-0.269)	0.675 (0.665-0.685)	0.176 (0.163-0.190)
Female patients	<i>0.785</i> (0.772-0.797)	<i>0.214</i> (0.194-0.236)	0.676 (0.662-0.692)	0.137 (0.123-0.155)	0.766 (0.754-0.779)	0.194 (0.176-0.216)	0.704 (0.689-0.718)	0.129 (0.116-0.145)
Age group (years)								
<40	<i>0.818</i> (0.797-0.837)	<i>0.222</i> (0.193-0.256)	0.739 (0.718-0.761)	0.153 (0.130-0.179)	0.804 (0.784-0.824)	0.213 (0.184-0.249)	0.738 (0.717-0.758)	0.120 (0.104-0.138)
40-59	<i>0.758</i> (0.744-0.772)	<i>0.251</i> (0.230-0.275)	0.609 (0.592-0.626)	0.159 (0.143-0.176)	0.734 (0.721-0.749)	0.226 (0.208-0.248)	0.640 (0.626-0.655)	0.149 (0.134-0.165)
≥60	<i>0.779</i> (0.768-0.790)	<i>0.258</i> (0.240-0.278)	0.663 (0.649-0.676)	0.196 (0.181-0.214)	0.757 (0.745-0.768)	0.235 (0.218-0.254)	0.700 (0.689-0.712)	0.177 (0.162-0.192)

^aXGBoost: extreme gradient boosting.

^bNEWS: National Early Warning Score.

^cThe best results in each subgroup are italicized.

Figure 3. Performance of the models on the overall test set across the different prediction time windows. We evaluated the performance of each model for deterioration prediction within 6, 12, 24, and 36 hours. AUPRC: area under the precision-recall curve; AUROC: area under the receiver operating characteristic curve; NEWS: National Early Warning Score; NN: neural network; XGBoost: extreme gradient boosting.



Performance Across Different Patient Subgroups

In Table 3, we have also summarized the performance of the models across different patient subgroups within the test set. Across the male population, the XGBoost model achieved the best performance in terms of AUROC values (0.775, 95% CI 0.764-0.784), with an improvement of 0.099 (95% CI 0.090-0.109; $P < .001$) compared with the NEWS system. The XGBoost model also achieved the best performance in the female population with an AUROC value of 0.785 (95% CI 0.772-0.797) and an AUPRC value of 0.214 (95% CI 0.194-0.236), which corresponds to improvements of 0.081 (95% CI 0.070-0.092; $P < .001$) in AUROC value and 0.084 (95% CI 0.070-0.099; $P < .001$) in AUPRC value compared with the NEWS system.

In the different age subpopulations, the XGBoost model achieved the best results (AUROC 0.758-0.818), followed by the NN model (AUROC 0.721-0.760). In the population consisting of patients aged <40 years, the XGBoost model achieved the best performance in terms of AUROC value (0.818, 95% CI 0.797-0.837) and AUPRC value (0.222, 95% CI 0.193-0.256), with improvements of 0.080 (95% CI 0.064-0.097; $P < .001$) and 0.102 (95% CI 0.082-0.125; $P < .001$) in AUROC and AUPRC values, respectively, compared with the NEWS system. In the group consisting of patients aged 40-59 years, the XGBoost model achieved an AUROC value of 0.758 (95% CI 0.744-0.772) and an AUPRC value of 0.251 (95% CI

0.230-0.275), with improvements of 0.118 (95% CI 0.104-0.133; $P < .001$) and 0.102 (95% CI 0.087-0.119; $P < .001$) in AUROC and AUPRC values, respectively, compared with the NEWS system.

Finally, in the group consisting of patients aged ≥ 60 years, the XGBoost model achieved an AUROC value of 0.779 (95% CI 0.768-0.790) and an AUPRC value of 0.258 (95% CI 0.240-0.278), with improvements of 0.078 (95% CI 0.069-0.088; $P < .001$) and 0.081 (95% CI 0.068-0.094; $P < .001$) in AUROC and AUPRC values, respectively, compared with the NEWS system.

Performance Based on a Temporal Data Split

Given that most of the patient cohort during 2020-2021 consisted of patients with COVID-19 infection, we investigated the impact of increasing the size of the training set based on a temporal data split for deterioration within 24 hours. To do so, we defined four training sets that encompassed data collected during (1) 2016, (2) 2016-2017, (3) 2016-2018, and (4) 2016-2019. We defined a new test set that included the observation sets of all patients admitted to the hospital in 2020. We excluded any data collected during 2015 and 2021 because our data set only included a few months from both years. The new test set consisted of 517 unique patients (307/517, 59.4% male patients), with 638 encounters associated with an average age of 54.0 years and 23,227 observation sets (1208/23,227, 5.2% deterioration within 24 hours).

We summarize the results in [Table 4](#). We observed that increasing the size of the training set yielded marginal improvements in terms of AUROC and AUPRC values across all models. The NN model saw the largest improvement in AUROC value, which increased from 0.706 (95% CI

0.688-0.722) to 0.754 (95% CI 0.739-0.769), whereas the XGBoost model saw the largest improvement in AUPRC value, which increased from 0.207 (0.187-0.229) to 0.250 (95% CI 0.226-0.276).

Table 4. Model performance based on a temporal data split for deterioration within 24 hours. We performed a temporal data split for the training and test sets. We fixed the test set to patient encounters recorded during 2020, whereas we expanded the training set gradually to eventually include encounters recorded between 2016 and 2019. We report area under the receiver operating characteristic curve (AUROC) and area under the precision-recall curve (AUPRC) values with 95% CIs.

Training set	Deterioration within 24 h, n (%)	XGBoost ^a		Logistic regression		Neural network	
		AUROC (95% CI)	AUPRC (95% CI)	AUROC (95% CI)	AUPRC (95% CI)	AUROC (95% CI)	AUPRC (95% CI)
2016: n=68,499	2788 (4.1)	0.740 (0.724-0.754)	0.207 (0.187-0.229)	0.679 (0.660-0.696)	0.204 (0.182-0.228)	0.706 (0.688-0.722)	0.211 (0.188-0.233)
2016-2017: n=159,888	7062 (4.4)	0.763 (0.747-0.778)	0.232 (0.211-0.257)	0.687 (0.669-0.704)	0.213 (0.191-0.237)	0.744 (0.727-0.759)	0.241 (0.216-0.266)
2016-2018: n=285,733	12,352 (4.3)	0.758 (0.742-0.773)	0.242 (0.218-0.267)	0.69 (0.671-0.706)	0.216 (0.194-0.240)	0.745 (0.728-0.760)	0.237 (0.213-0.262)
2016-2019: n=431,503	19,261 (4.5)	0.778 (0.763-0.792)	0.25 (0.226-0.276)	0.688 (0.670-0.705)	0.215 (0.192-0.239)	0.754 (0.739-0.769)	0.233 (0.211-0.259)

^aXGBoost: extreme gradient boosting.

Interpretability Results

[Table 5](#) shows the overall importance of each input feature in the XGBoost model predicting deterioration within 24 hours. The plots for the other time windows (6, 12, and 36 hours) are shown in [Figure S1](#) in [Multimedia Appendix 2](#). We observed a similar pattern across all time window values. We noted that temperature is the most important feature, followed closely by respiratory rate, systolic blood pressure, heart rate, level of consciousness, oxygen saturation, and finally provision of supplementary oxygen.

[Figure 4](#) shows the SHAP values and the corresponding feature values of the observation sets that were assigned the highest and lowest predictions of deterioration. We observed that for all observation sets with the highest assigned probabilities, the factors contributing the most were high or low systolic blood pressure measurements, level of consciousness where a value of 3 indicated that the patient was unconscious, and high heart rate measurements. In the 5 observation sets with the lowest assigned probabilities, the patients displayed mostly normal vital signs measurements.

[Figure 5](#) shows the SHAP partial dependence plots for 6 (86%) of the 7 input features. We observed that for continuous

variables (eg, heart rate, respiratory rate, oxygen saturation, temperature, and systolic blood pressure), there is a range of values for which the SHAP contributions are the lowest; for example, for heart rate, the average SHAP value encounters the sharpest drop between approximately 50 and 100 beats per minute. For oxygen saturation, we observed that the SHAP values decreased as oxygen saturation increased to >80%, whereas for respiratory rate, we observed that SHAP values increased as respiratory rate increased. Level of consciousness is a binary variable, and it can be observed in [Figure 5](#) that the average SHAP value for level of consciousness varies based on whether the patient is conscious.

[Table 6](#) shows the Pearson correlation coefficients and Spearman rank correlation coefficients between the SHAP values and the feature values, as well as the LR coefficients and odds ratios. We observed that level of consciousness shows the highest level of correlation (Pearson correlation coefficient=0.950, Spearman rank correlation coefficient=1.000, and LR coefficient=0.532). We also noted that the LR coefficients are aligned with those of SHAP, based on the relative ranking of the features with the calculated Pearson coefficients and the LR coefficients. Temperature exhibits the lowest level of correlation, perhaps because of the complexity of the nonlinear relationship between the feature and the outcome variable.

Table 5. Feature importance of the extreme gradient boosting model. We present the results of our Shapley additive explanations (SHAP) analysis for the extreme gradient boosting model for the deterioration within each of our proposed time windows. We provide the mean of the absolute SHAP value for each of the 7 input features.

Vital sign	Mean of absolute SHAP value			
	36 hours	24 hours	12 hours	6 hours
Temperature	0.019	0.018	0.018	0.018
Respiratory rate	0.016	0.015	0.015	0.012
Systolic blood pressure	0.013	0.013	0.013	0.011
Heart rate	0.011	0.010	0.010	0.008
Level of consciousness	0.003	0.003	0.003	0.002
Oxygen saturation	0.003	0.003	0.003	0.002
Supplementary oxygen	0.000	0.000	0.000	0.000

Figure 4. Feature importance of the highest and lowest predictions of deterioration in the test set. We present the Shapley additive explanations (SHAP) values for (A) 5 observation sets with the highest predictions of deterioration assigned by the extreme gradient boosting (XGBoost) model in the test set and (B) 5 observation sets with the lowest predictions of deterioration. We confirmed that all observation sets in (A) did indeed experience an adverse event within 24 hours, whereas all observation sets in (B) did not. Note that temperature values are displayed in degrees Fahrenheit. For a higher-resolution version of this figure, see [Multimedia Appendix 3](#).

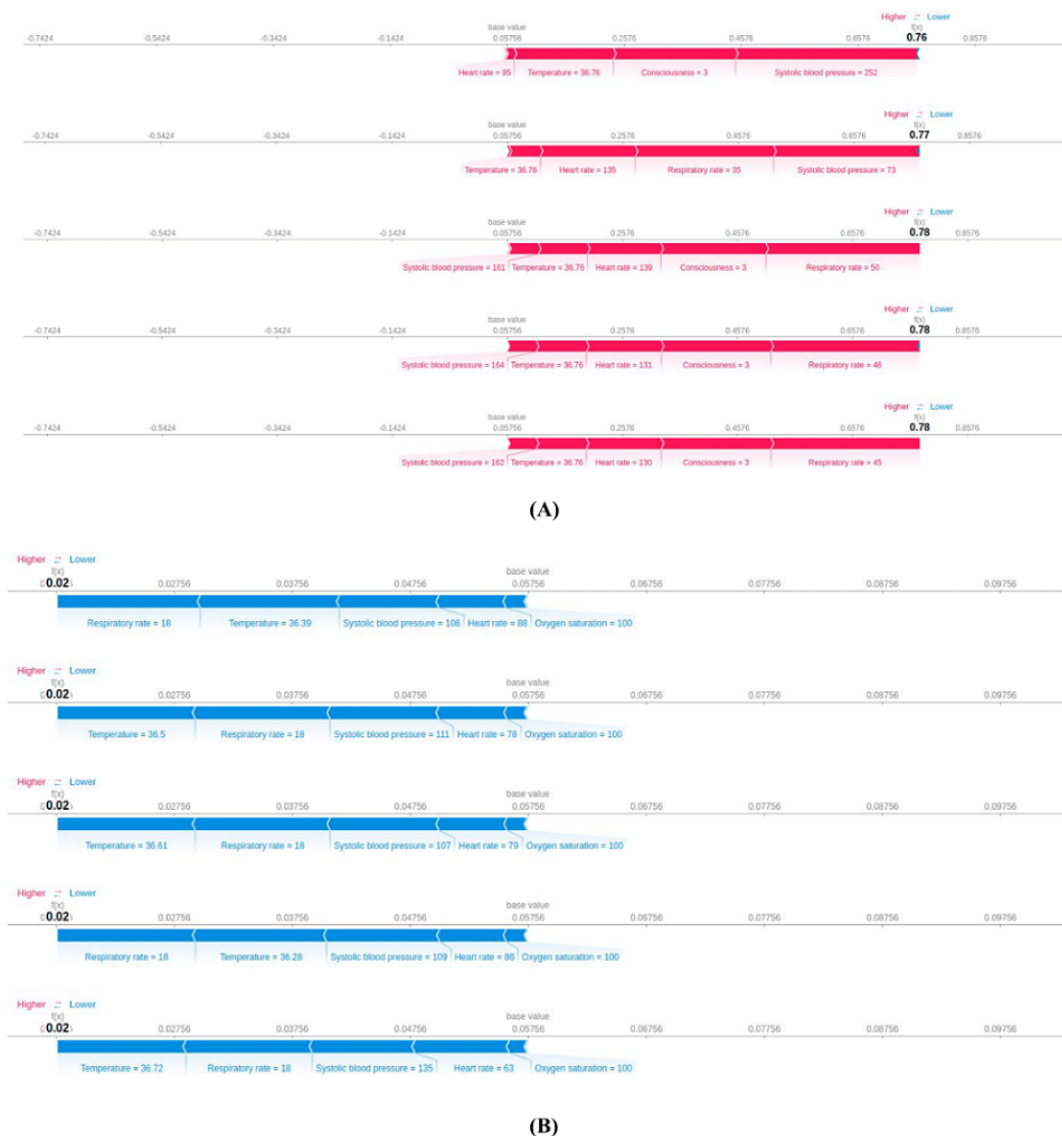


Figure 5. Partial dependence plots for the input features. We present the Shapley additive explanations (SHAP) partial dependence plots for six input features: (A) heart rate, (B) respiratory rate, (C) oxygen saturation, (D) temperature, (E) level of consciousness, and (F) systolic blood pressure. The partial dependence plot for supplementary oxygen is a flat line; hence, it has been omitted from the figure.

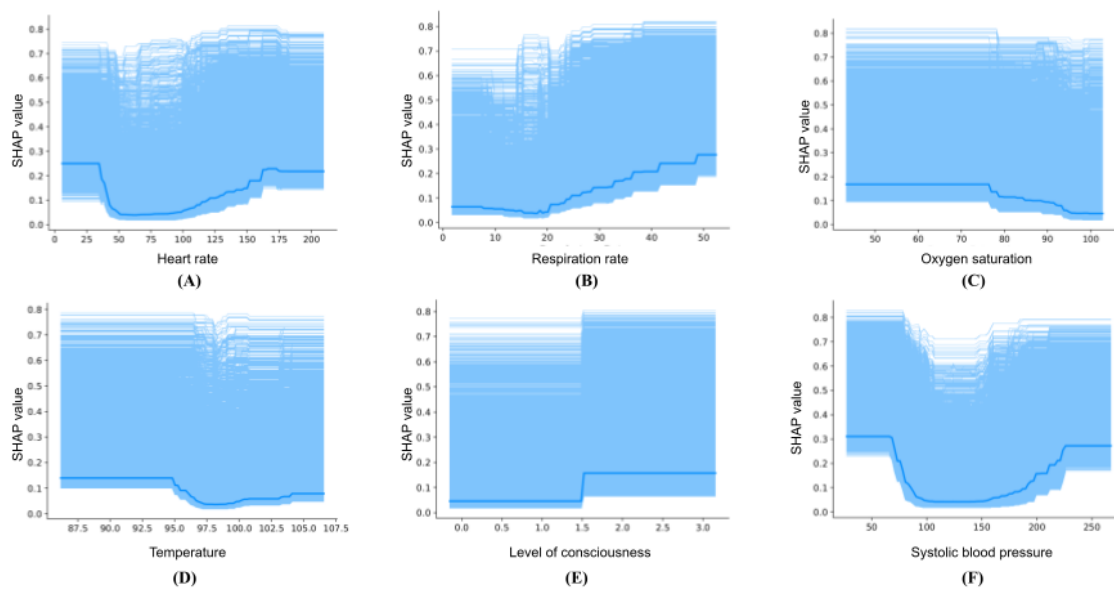


Table 6. Feature correlation values. We summarize the Pearson correlation coefficients and Spearman rank correlation coefficients between the calculated Shapley additive explanations (SHAP) values and the respective features. For each feature, we also present the coefficients of the logistic regression model and their respective odds ratios.

	Pearson correlation coefficient	Spearman rank correlation coefficient	Logistic regression coefficient	Logistic regression odds ratio
Heart rate	0.606	0.816	0.016	1.016
Respiratory rate	0.792	0.417	0.128	1.137
Oxygen saturation	-0.713	-0.694	-0.073	0.930
Temperature	0.006	-0.068	0.001	1.002
Level of consciousness	0.950	1.000	0.532	1.702
Supplementary oxygen	0.021	0.080	0.011	1.011
Systolic blood pressure	-0.060	-0.099	0.001	1.001

Discussion

Principal Findings

EWS systems provide a standardized method for the detection of patient deterioration [46]. Despite the proliferation of EWS systems in electronic health record systems, they are often developed based on heuristics or data acquired from a specific patient cohort [12]. One such EWS system is the NEWS system [4], which is recommended by the Royal College of Physicians and is currently in use in some hospitals in the United Arab Emirates. In this work, we developed and evaluated data-driven deterioration prediction models using ML and real-world data collected at a local hospital. We compared the performance of the ML models with that of the NEWS system in a holdout test set consisting of 2548 encounters and 95,755 observation sets in terms of AUROC and AUPRC values.

Our study has several strengths. First, in the overall population, our results showed that the XGBoost model and the NN model achieved the best performance with improvements of 0.096 (95% CI 0.088-0.103; $P < .001$) and 0.074 (95% CI 0.067-0.081;

$P < .001$), respectively, compared with the NEWS system. This is consistent with the findings of other studies, where the XGBoost model predominantly achieved the best performance compared with other models, especially with tabular input data [34,47,48]. Considering the performance improvement with respect to the NEWS system, we suggest in this case that a hospital is likely to benefit more by developing its own models using cohort-specific data, instead of relying on external models [49]. However, this requires expertise and computational resources that may not always be readily available. In addition, we showed that although the models' performance remained stable as the training sets were expanded, and more data were collected, future work should focus on tackling distribution shifts owing to changes in practice over time or changes in patient phenotype and demographics. The discrepancy in performance across all models when using a random data split compared with a temporal split also highlights the importance of choosing training and test sets that best reflect the eligible population during model deployment and implementation.

Another strength of our study is that we assessed the performance of the models across different deterioration

windows. We showed that as the prediction window increased in size, the predictive performance of all models decreased because the level of difficulty of the prediction tasks increased. This implies that, in practice, one must deploy the model that best aligns with the interventions that can be implemented. We also assessed the importance of the input features as an interpretability mechanism. In predicting deterioration within 24 hours, respiratory rate was among the top 2 most important features. This is in line with existing work that emphasizes the importance of respiratory rate as a clinical biomarker and indicator of patient status [50].

Despite the contributions of our study in proposing a new deterioration prediction model for the United Arab Emirates population, our study has some limitations. We only assessed the performance of our model using an internal test set from a single center because we did not have access to any external validation cohorts. In addition, our model relied on a small set of 7 input features, mostly vital signs, and we did not include any other variables that may be indicative of deterioration, such as laboratory test results. We performed a patient-level split across the training, validation, and test splits to avoid data leakage across the data splits. However, this could potentially bias the learning of the model owing to patients having multiple encounters or observation sets within a specific data split. On average, each unique patient had 1.6, 1.5, and 1.5 encounters in the training, validation, and test sets, respectively; therefore, we suspect low levels of bias, although this is a limitation of the training strategy. As we developed models that computed predictions every time an observation set was recorded, to mimic

EWS systems in real time, we also included all observation sets of all encounters. In future work, more advanced data-split training and evaluation strategies can be investigated for encounter-level predictions with more advanced methods that consider time-series analysis.

Future work should also focus on the development of multimodal EWS systems, including imaging modalities such as chest x-ray images [51]. However, this depends on the target population of the EWS system and the availability of multimodal data. We also did not assess the performance of the latest version of the NEWS system [1,52], also referred to as NEWS2, which introduced specific alerting thresholds for patients with hypercapnic respiratory failure in a current or previous encounter, and this is an area of future work. Another area of future work with expected clinical impact would be to study how existing patient management protocols can be re-evaluated with respect to the model's predictions and marginal risk measures computed using SHAP analysis for the input features.

Conclusions

In conclusion, we developed and evaluated deterioration prediction models using ML and a real-world data set and compared their performance with that of the NEWS system, which is commonly used in practice. In future work, we will seek to evaluate the performance of the XGBoost model in a silent prospective validation study to verify further areas of improvement. Although we developed models specific to our patient cohort, we believe that our framework may be useful to other researchers interested in developing and evaluating deterioration prediction models.

Acknowledgments

The authors would like to thank the High Performance Computing team at New York University Abu Dhabi (NYUAD) and Helen Sun at Cleveland Clinic Abu Dhabi for their support. FES received funding from the NYUAD Center for Artificial Intelligence and Robotics, funded by Tamkeen under the NYUAD Research Institute Award (CG010). The funders had no role in the study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Data Availability

Synthetic data that are randomly generated to closely resemble the original data set used in this study is available in [Multimedia Appendix 4](#)

Conflicts of Interest

None declared.

Multimedia Appendix 1

Transparent Reporting of a Multivariable Prediction Model for Individual Prognosis or Diagnosis (TRIPOD) checklist.

[\[DOCX File, 486 KB - ai_v2i1e45257_app1.docx \]](#)

Multimedia Appendix 2

Supplementary figures and tables.

[\[DOCX File, 93 KB - ai_v2i1e45257_app2.docx \]](#)

Multimedia Appendix 3

Feature importance of the highest and lowest predictions of deterioration in the test set. We present the Shapley additive explanations (SHAP) values for (A) 5 observation sets with the highest predictions of deterioration assigned by the extreme gradient boosting

(XGBoost) model in the test set and (B) 5 observation sets with the lowest predictions of deterioration. We confirmed that all observation sets in (A) did indeed experience an adverse event within 24 hours, whereas all observation sets in (B) did not. Note that temperature values are displayed in degrees Fahrenheit.

[DOCX File, 222 KB - [ai_v2i1e45257_app3.docx](#)]

Multimedia Appendix 4

Synthetic data that are randomly generated to closely resemble the original data set used in this study.

[ZIP File (Zip Archive), 78 KB - [ai_v2i1e45257_app4.zip](#)]

References

1. Prytherch DR, Smith GB, Schmidt PE, Featherstone PI. ViEWS--towards a national early warning score for detecting adult inpatient deterioration. *Resuscitation* 2010 Aug;81(8):932-937. [doi: [10.1016/j.resuscitation.2010.04.014](#)] [Medline: [20637974](#)]
2. Shamout F, Zhu T, Clifton L, Briggs J, Prytherch D, Meredith P, et al. Early warning score adjusted for age to predict the composite outcome of mortality, cardiac arrest or unplanned intensive care unit admission using observational vital-sign data: a multicentre development and validation. *BMJ Open* 2019 Nov 19;9(11):e033301 [FREE Full text] [doi: [10.1136/bmjopen-2019-033301](#)] [Medline: [31748313](#)]
3. Subbe CP, Kruger M, Rutherford P, Gemmel L. Validation of a modified early warning score in medical admissions. *QJM* 2001 Oct;94(10):521-526. [doi: [10.1093/qjmed/94.10.521](#)] [Medline: [11588210](#)]
4. National early warning score (news): standardising the assessment of acute-illness severity in the NHS. Royal College of Physicians. 2012. URL: <https://www.rcplondon.ac.uk/projects/outputs/national-early-warning-score-news-2> [accessed 2023-08-24]
5. Fu LH, Schwartz J, Moy A, Knaplund C, Kang MJ, Schnock KO, et al. Development and validation of early warning score system: a systematic literature review. *J Biomed Inform* 2020 May;105:103410 [FREE Full text] [doi: [10.1016/j.jbi.2020.103410](#)] [Medline: [32278089](#)]
6. Kaul M, Snethen J, Kelber ST, Zimmanck K, Maletta K, Meyer M. Implementation of the bedside paediatric early warning system (BedsidePEWS) for nurse identification of deteriorating patients. *J Spec Pediatr Nurs* 2014 Oct;19(4):339-349. [doi: [10.1111/jspn.12092](#)] [Medline: [25348360](#)]
7. Zhang Y, Zhuo H, Bi H, Wu J. Feasibility investigation of developing a revised MEWS score for cardiovascular specialty. *Am J Nurs Sci* 2021 Jun;10(3):169-173 [FREE Full text] [doi: [10.11648/j.ajns.20211003.16](#)]
8. Downey CL, Tahir W, Randell R, Brown JM, Jayne DG. Strengths and limitations of early warning scores: a systematic review and narrative synthesis. *Int J Nurs Stud* 2017 Nov;76:106-119 [FREE Full text] [doi: [10.1016/j.ijnurstu.2017.09.003](#)] [Medline: [28950188](#)]
9. Finnikin S, Wilke V. What's behind the NEWS? National early warning scores in primary care. *Br J Gen Pract* 2020 May 28;70(695):272-273 [FREE Full text] [doi: [10.3399/bjgp20X709361](#)] [Medline: [32269041](#)]
10. Opio MO, Nansubuga G, Kellett J. Validation of the VitalPAC™ Early Warning Score (ViEWS) in acutely ill medical patients attending a resource-poor hospital in sub-Saharan Africa. *Resuscitation* 2013 Jun;84(6):743-746. [doi: [10.1016/j.resuscitation.2013.02.007](#)] [Medline: [23438452](#)]
11. Baker T, Blixt J, Lugazia E, Schell CO, Mulungu M, Milton A, et al. Single deranged physiologic parameters are associated with mortality in a low-income country. *Crit Care Med* 2015 Oct;43(10):2171-2179. [doi: [10.1097/CCM.0000000000001194](#)] [Medline: [26154933](#)]
12. Youssef A, Kouchaki S, Shamout F, Armstrong J, El-Bouri R, Taylor T, et al. Development and validation of early warning score systems for COVID-19 patients. *Healthc Technol Lett* 2021 Oct 27;8(5):105-117 [FREE Full text] [doi: [10.1049/htl2.12009](#)] [Medline: [34221413](#)]
13. Henry KE, Hager DN, Pronovost PJ, Saria S. A targeted real-time early warning score (TREWScore) for septic shock. *Sci Transl Med* 2015 Aug 05;7(299):299ra122. [doi: [10.1126/scitranslmed.aab3719](#)] [Medline: [26246167](#)]
14. Garcia-Gutiérrez S, Esteban-Aizpiri C, Lafuente I, Barrio I, Quiros R, Quintana JM, COVID-REDISSEC Working Group. Author correction: machine learning-based model for prediction of clinical deterioration in hospitalized patients by COVID 19. *Sci Rep* 2022 May 12;12(1):7811 [FREE Full text] [doi: [10.1038/s41598-022-12247-9](#)] [Medline: [35552505](#)]
15. Nistal-Nuño B. Developing machine learning models for prediction of mortality in the medical intensive care unit. *Comput Methods Programs Biomed* 2022 Apr;216:106663. [doi: [10.1016/j.cmpb.2022.106663](#)] [Medline: [35123348](#)]
16. Masum S, Hopgood A, Stefan S, Flashman K, Khan J. Data analytics and artificial intelligence in predicting length of stay, readmission, and mortality: a population-based study of surgical management of colorectal cancer. *Discov Oncol* 2022 Feb 28;13(1):11 [FREE Full text] [doi: [10.1007/s12672-022-00472-7](#)] [Medline: [35226196](#)]
17. Gopukumar D, Ghoshal A, Zhao H. Predicting readmission charges billed by hospitals: machine learning approach. *JMIR Med Inform* 2022 Aug 30;10(8):e37578 [FREE Full text] [doi: [10.2196/37578](#)] [Medline: [35896038](#)]

18. Wang Z, Chen X, Tan X, Yang L, Kannapur K, Vincent JL, et al. Using deep learning to identify high-risk patients with heart failure with reduced ejection fraction. *J Health Econ Outcomes Res* 2021 Jul 29;8(2):6-13 [FREE Full text] [doi: [10.36469/jheor.2021.25753](https://doi.org/10.36469/jheor.2021.25753)] [Medline: [34414250](https://pubmed.ncbi.nlm.nih.gov/34414250/)]
19. Ong ME, Lee Ng CH, Goh K, Liu N, Koh ZX, Shahidah N, et al. Prediction of cardiac arrest in critically ill patients presenting to the emergency department using a machine learning score incorporating heart rate variability compared with the modified early warning score. *Crit Care* 2012 Jun 21;16(3):R108 [FREE Full text] [doi: [10.1186/cc11396](https://doi.org/10.1186/cc11396)] [Medline: [22715923](https://pubmed.ncbi.nlm.nih.gov/22715923/)]
20. Desautels T, Calvert J, Hoffman J, Jay M, Kerem Y, Shieh L, et al. Prediction of sepsis in the intensive care unit with minimal electronic health record data: a machine learning approach. *JMIR Med Inform* 2016 Sep 30;4(3):e28 [FREE Full text] [doi: [10.2196/medinform.5909](https://doi.org/10.2196/medinform.5909)] [Medline: [27694098](https://pubmed.ncbi.nlm.nih.gov/27694098/)]
21. Lee YJ, Cho KJ, Kwon O, Park H, Lee Y, Kwon JM, et al. A multicentre validation study of the deep learning-based early warning score for predicting in-hospital cardiac arrest in patients admitted to general wards. *Resuscitation* 2021 Apr 22;163:78-85 [FREE Full text] [doi: [10.1016/j.resuscitation.2021.04.013](https://doi.org/10.1016/j.resuscitation.2021.04.013)] [Medline: [33895236](https://pubmed.ncbi.nlm.nih.gov/33895236/)]
22. Bolourani S, Brenner M, Wang P, McGinn T, Hirsch JS, Barnaby D, Northwell COVID-19 Research Consortium. A machine learning prediction model of respiratory failure within 48 hours of patient admission for COVID-19: model development and validation. *J Med Internet Res* 2021 Feb 10;23(2):e24246 [FREE Full text] [doi: [10.2196/24246](https://doi.org/10.2196/24246)] [Medline: [33476281](https://pubmed.ncbi.nlm.nih.gov/33476281/)]
23. Shamout FE, Zhu T, Sharma P, Watkinson PJ, Clifton DA. Deep interpretable early warning system for the detection of clinical deterioration. *IEEE J Biomed Health Inform* 2020 Feb;24(2):437-446. [doi: [10.1109/JBHI.2019.2937803](https://doi.org/10.1109/JBHI.2019.2937803)] [Medline: [31545746](https://pubmed.ncbi.nlm.nih.gov/31545746/)]
24. Trevisi G, Caccavella VM, Scerrati A, Signorelli F, Salamone GG, Orsini K, et al. Machine learning model prediction of 6-month functional outcome in elderly patients with intracerebral hemorrhage. *Neurosurg Rev* 2022 Aug;45(4):2857-2867 [FREE Full text] [doi: [10.1007/s10143-022-01802-7](https://doi.org/10.1007/s10143-022-01802-7)] [Medline: [35522333](https://pubmed.ncbi.nlm.nih.gov/35522333/)]
25. Min X, Yu B, Wang F. Predictive modeling of the hospital readmission risk from patients' claims data using machine learning: a case study on COPD. *Sci Rep* 2019 Feb 20;9(1):2362 [FREE Full text] [doi: [10.1038/s41598-019-39071-y](https://doi.org/10.1038/s41598-019-39071-y)] [Medline: [30787351](https://pubmed.ncbi.nlm.nih.gov/30787351/)]
26. Mathis MR, Engoren MC, Williams AM, Biesterveld BE, Croteau AJ, Cai L, BCIL Collaborators Group. Prediction of postoperative deterioration in cardiac surgery patients using electronic health record and physiologic waveform data. *Anesthesiology* 2022 Nov 01;137(5):586-601 [FREE Full text] [doi: [10.1097/ALN.0000000000004345](https://doi.org/10.1097/ALN.0000000000004345)] [Medline: [35950802](https://pubmed.ncbi.nlm.nih.gov/35950802/)]
27. Wu KH, Cheng FJ, Tai HL, Wang JC, Huang YT, Su CM, et al. Predicting in-hospital mortality in adult non-traumatic emergency department patients: a retrospective comparison of the Modified Early Warning Score (MEWS) and machine learning approach. *PeerJ* 2021 Aug 24;9:e11988 [FREE Full text] [doi: [10.7717/peerj.11988](https://doi.org/10.7717/peerj.11988)] [Medline: [34513328](https://pubmed.ncbi.nlm.nih.gov/34513328/)]
28. Noy O, Coster D, Metzger M, Atar I, Shenhar-Tsarfaty S, Berliner S, et al. A machine learning model for predicting deterioration of COVID-19 inpatients. *Sci Rep* 2022 Feb 16;12(1):2630 [FREE Full text] [doi: [10.1038/s41598-022-05822-7](https://doi.org/10.1038/s41598-022-05822-7)] [Medline: [35173197](https://pubmed.ncbi.nlm.nih.gov/35173197/)]
29. Jakob CE, Mahajan UM, Oswald M, Stecher M, Schons M, Mayerle J, LEOSS Study group. Prediction of COVID-19 deterioration in high-risk patients at diagnosis: an early warning score for advanced COVID-19 developed by machine learning. *Infection* 2022 Apr;50(2):359-370 [FREE Full text] [doi: [10.1007/s15010-021-01656-z](https://doi.org/10.1007/s15010-021-01656-z)] [Medline: [34279815](https://pubmed.ncbi.nlm.nih.gov/34279815/)]
30. Shin S, Austin PC, Ross HJ, Abdel-Qadir H, Freitas C, Tomlinson G, et al. Machine learning vs. conventional statistical models for predicting heart failure readmission and mortality. *ESC Heart Fail* 2021 Feb;8(1):106-115 [FREE Full text] [doi: [10.1002/ehf2.13073](https://doi.org/10.1002/ehf2.13073)] [Medline: [33205591](https://pubmed.ncbi.nlm.nih.gov/33205591/)]
31. Collins GS, Reitsma JB, Altman DG, Moons KG. Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis (TRIPOD): the TRIPOD Statement. *Br J Surg* 2015 Feb;102(3):148-158 [FREE Full text] [doi: [10.1002/bjs.9736](https://doi.org/10.1002/bjs.9736)] [Medline: [25627261](https://pubmed.ncbi.nlm.nih.gov/25627261/)]
32. Fuentes S, Chowdhury YS. Fraction of inspired oxygen. *StatPearls*. 2021. URL: <https://www.ncbi.nlm.nih.gov/books/NBK560867/> [accessed 2023-08-24]
33. Gerry S, Bonnici T, Birks J, Kirtley S, Virdee PS, Watkinson PJ, et al. Early warning scores for detecting deterioration in adult hospital patients: systematic review and critical appraisal of methodology. *BMJ* 2020 May 20;369:m1501 [FREE Full text] [doi: [10.1136/bmj.m1501](https://doi.org/10.1136/bmj.m1501)] [Medline: [32434791](https://pubmed.ncbi.nlm.nih.gov/32434791/)]
34. Chen T, Guestrin CE. XGBoost: a scalable tree boosting system. In: *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. 2016 Presented at: KDD '16; August 13-17, 2016; San Francisco, CA p. 785-794 URL: <https://dl.acm.org/doi/10.1145/2939672.2939785> [doi: [10.1145/2939672.2939785](https://doi.org/10.1145/2939672.2939785)]
35. Goldhill DR, McNarry AF. Physiological abnormalities in early warning scores are related to mortality in adult inpatients. *Br J Anaesth* 2004 Jun;92(6):882-884 [FREE Full text] [doi: [10.1093/bja/ae113](https://doi.org/10.1093/bja/ae113)] [Medline: [15064245](https://pubmed.ncbi.nlm.nih.gov/15064245/)]
36. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, et al. Machine learning for neuroimaging with scikit-learn. *J Mach Learn Res* 2011;12(85):2825-2830 [FREE Full text] [doi: [10.3389/fninf.2014.00014](https://doi.org/10.3389/fninf.2014.00014)]
37. Jang DH, Kim J, Jo YH, Lee JH, Hwang JE, Park SM, et al. Developing neural network models for early detection of cardiac arrest in emergency department. *Am J Emerg Med* 2020 Jan;38(1):43-49. [doi: [10.1016/j.ajem.2019.04.006](https://doi.org/10.1016/j.ajem.2019.04.006)] [Medline: [30982559](https://pubmed.ncbi.nlm.nih.gov/30982559/)]

38. Klambauer G, Unterthiner T, Mayr A, Hochreiter S. Self-normalizing neural networks. In: Proceedings of the 31st International Conference on Neural Information Processing Systems. 2017 Presented at: NIPS '17; December 4-9, 2017; Long Beach, CA p. 972-981 URL: <https://dl.acm.org/doi/10.5555/3294771.3294864>
39. Paszke A, Gross S, Massa F, Lerer A, Bradbury J, Chanan G, et al. Pytorch: An imperative style, high-performance deep learning library. In: Proceedings of the 33rd International Conference on Neural Information Processing Systems. 2019 Presented at: NIPS '19; December 8-14, 2019; Vancouver, BC p. 8026-8037 URL: <https://dl.acm.org/doi/10.5555/3454287.3455008>
40. DiCiccio TJ, Efron B. Bootstrap confidence intervals. *Statist Sci* 1996 Aug;11(3):189-228 [FREE Full text] [doi: [10.1214/ss/1032280214](https://doi.org/10.1214/ss/1032280214)]
41. Chihara LM, Hesterberg TC. *Mathematical Statistics with Resampling and R*. Hoboken, NJ: John Wiley & Sons; 2022.
42. James B, Bengio Y. Random search for hyper-parameter optimization. *J Mach Learn Res* 2012 Feb 01;13(2):281-305 [FREE Full text]
43. Zhu W, Razavian N. Variationally regularized graph-based representation learning for electronic health records. In: Proceedings of the 2021 Conference on Health, Inference, and Learning. 2021 Presented at: CHIL '21; April 8-10, 2021; Virtual Event p. 1-13 URL: <https://dl.acm.org/doi/10.1145/3450439.3451855> [doi: [10.1145/3450439.3451855](https://doi.org/10.1145/3450439.3451855)]
44. 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 Mar 04;10(3):e0118432 [FREE Full text] [doi: [10.1371/journal.pone.0118432](https://doi.org/10.1371/journal.pone.0118432)] [Medline: [25738806](https://pubmed.ncbi.nlm.nih.gov/25738806/)]
45. Lee S, Lundberg SM. A unified approach to interpreting model predictions. In: Proceedings of the 31st International Conference on Neural Information Processing Systems. 2017 Presented at: NIPS '17; December 4-9, 2017; Long Beach, CA p. 4774-4777 URL: <https://dl.acm.org/doi/10.5555/3295222.3295230>
46. Morgan RJ, Wright MM. In defence of early warning scores. *Br J Anaesth* 2007 Nov;99(5):747-748 [FREE Full text] [doi: [10.1093/bja/aem286](https://doi.org/10.1093/bja/aem286)] [Medline: [17933804](https://pubmed.ncbi.nlm.nih.gov/17933804/)]
47. Kobylarz Ribeiro J, dos Santos HD, Barletta F, da Silva MC, Vieira R, Morales HM, et al. A machine learning early warning system: multicenter validation in brazilian hospitals. In: Proceedings of the 33rd International Symposium on Computer-Based Medical Systems. 2020 Presented at: CBMS '20; July 28-30, 2020; Rochester, MN p. 321-326 URL: <https://ieeexplore.ieee.org/document/9183044> [doi: [10.1109/cbms49503.2020.00067](https://doi.org/10.1109/cbms49503.2020.00067)]
48. Wu TT, Lin XQ, Mu Y, Li H, Guo YS. Machine learning for early prediction of in-hospital cardiac arrest in patients with acute coronary syndromes. *Clin Cardiol* 2021 Feb 14;44(3):349-356. [doi: [10.1002/clc.23541](https://doi.org/10.1002/clc.23541)] [Medline: [33586214](https://pubmed.ncbi.nlm.nih.gov/33586214/)]
49. Futoma J, Simons M, Panch T, Doshi-Velez F, Celi LA. The myth of generalisability in clinical research and machine learning in health care. *The Lancet Digital Health* 2020 Sep;2(9):e489-e492. [doi: [10.1016/s2589-7500\(20\)30186-2](https://doi.org/10.1016/s2589-7500(20)30186-2)]
50. Mochizuki K, Shintani R, Mori K, Sato T, Sakaguchi O, Takeshige K, et al. Importance of respiratory rate for the prediction of clinical deterioration after emergency department discharge: a single-center, case-control study. *Acute Med Surg* 2016 Nov 10;4(2):172-178 [FREE Full text] [doi: [10.1002/ams2.252](https://doi.org/10.1002/ams2.252)] [Medline: [29123857](https://pubmed.ncbi.nlm.nih.gov/29123857/)]
51. Shamout FE, Shen Y, Wu N, Kaku A, Park J, Makino T, et al. An artificial intelligence system for predicting the deterioration of COVID-19 patients in the emergency department. *NPJ Digit Med* 2021 May 12;4(1):80 [FREE Full text] [doi: [10.1038/s41746-021-00453-0](https://doi.org/10.1038/s41746-021-00453-0)] [Medline: [33980980](https://pubmed.ncbi.nlm.nih.gov/33980980/)]
52. Smith GB, Redfern OC, Pimentel MA, Gerry S, Collins GS, Malycha J, et al. The National Early Warning Score 2 (NEWS2). *Clin Med (Lond)* 2019 May 15;19(3):260 [FREE Full text] [doi: [10.7861/clinmedicine.19-3-260](https://doi.org/10.7861/clinmedicine.19-3-260)] [Medline: [31092526](https://pubmed.ncbi.nlm.nih.gov/31092526/)]

Abbreviations

AUPRC: area under the precision-recall curve

AUROC: area under the receiver operating characteristic curve

EWS: early warning score

ICU: intensive care unit

LR: logistic regression

ML: machine learning

NEWS: National Early Warning Score

NN: neural network

SHAP: Shapley additive explanations

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 02.02.23; peer-reviewed by B Ru, KF Chen; comments to author 19.04.23; revised version received 19.06.23; accepted 01.08.23; published 06.11.23.

Please cite as:

Lashen H, St John TL, Almallah YZ, Sasidhar M, Shamout FE

Machine Learning Models Versus the National Early Warning Score System for Predicting Deterioration: Retrospective Cohort Study in the United Arab Emirates

JMIR AI 2023;2:e45257

URL: <https://ai.jmir.org/2023/1/e45257>

doi: [10.2196/45257](https://doi.org/10.2196/45257)

PMID: [38875543](https://pubmed.ncbi.nlm.nih.gov/38875543/)

©Hazem Lashen, Terrence Lee St John, Y Zaki Almallah, Madhu Sasidhar, Farah E Shamout. Originally published in JMIR AI (<https://ai.jmir.org>), 06.11.2023. This is an open-access article distributed under the terms of the Creative Commons Attribution License (<https://creativecommons.org/licenses/by/4.0/>), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR AI, is properly cited. The complete bibliographic information, a link to the original publication on <https://www.ai.jmir.org/>, as well as this copyright and license information must be included.

Original Paper

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

Max Rollwage¹, BSc, MSc, MPhil, PhD; Johanna Habicht¹, MSci; Keno Juechems¹, BSc, MSc, PhD; Ben Carrington¹, BSc; Sruthi Viswanathan¹, BTech, MRes; Mona Stylianou², BA, PGDip; Tobias U Hauser^{1,3,4,5}, PhD; Ross Harper¹, MA, MRes, PhD

¹Limbic Limited, London, United Kingdom

²Everyturn Mental Health, Gosforth, United Kingdom

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

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

⁵German Center for Mental Health (DZPG), Tübingen, Germany

Corresponding Author:

Max Rollwage, BSc, MSc, MPhil, PhD

Limbic Limited

Kemp House

160 City Road

London, EC1V 2NX

United Kingdom

Phone: 44 07491263783

Email: max@limbic.ai

Related Article:

This is a corrected version. See correction statement: <https://ai.jmir.org/2024/1/e57869>

Abstract

Background: Most mental health care providers face the challenge of increased demand for psychotherapy in the absence of increased funding or staffing. To overcome this supply-demand imbalance, care providers must increase the efficiency of service delivery.

Objective: In this study, we examined whether artificial intelligence (AI)-enabled digital solutions can help mental health care practitioners to use their time more efficiently, and thus reduce strain on services and improve patient outcomes.

Methods: In this study, we focused on the use of an AI solution (Limbic Access) to support initial patient referral and clinical assessment within the UK's National Health Service. Data were collected from 9 Talking Therapies services across England, comprising 64,862 patients.

Results: We showed that the use of this AI solution improves clinical efficiency by reducing the time clinicians spend on mental health assessments. Furthermore, we found improved outcomes for patients using the AI solution in several key metrics, such as reduced wait times, reduced dropout rates, improved allocation to appropriate treatment pathways, and, most importantly, improved recovery rates. When investigating the mechanism by which the AI solution achieved these improvements, we found that the provision of clinically relevant information ahead of clinical assessment was critical for these observed effects.

Conclusions: Our results emphasize the utility of using AI solutions to support the mental health workforce, further highlighting the potential of AI solutions to increase the efficiency of care delivery and improve clinical outcomes for patients.

(JMIR AI 2023;2:e44358) doi:[10.2196/44358](https://doi.org/10.2196/44358)

KEYWORDS

artificial intelligence; National Health Service; NHS; Improving Access to Psychological Therapies; IAPT; mental health; mental health assessment; triage; decision-support; referral; chatbot; psychotherapy; conversational agent; assessment; Talking Therapies

Introduction

Background

Common mental illnesses have become the leading cause of disability worldwide [1]. Access to high-quality mental health care is therefore crucial, with up to 25% of the population experiencing depression or anxiety disorders [2,3]. The COVID-19 pandemic has further highlighted the need for accessible mental health treatment, precipitating increased cases of anxiety, depression, and other mental health symptoms [4-9]. Addressing this high demand is challenging for many mental health services that already struggle to provide adequate treatment with limited resources, resulting in impaired patient experience and, ultimately, worse treatment outcomes [10].

One particular challenge that mental health services face is the long wait time between the point from when a patient seeks support and when they begin treatment. For instance, in the English National Health Service (NHS), between 2021 and 2022, 31% of referrals to Talking Therapy services dropped off the wait list before starting treatment, and 9% of patients waited for >6 weeks for their clinical assessment [11]. In addition, a further 47% of patients experienced *hidden waits* of >28 days between clinical assessment and their first treatment session, contrary to the guidance from the National Institute of Health and Care Excellence, which highlights the importance of timely access to treatment [12].

Notably, against the backdrop of rising referrals, the needs of patients are unlikely to be addressed through an increase in the clinical workforce; in fact, there exists a national shortage of qualified staff [13]. To remedy this precarious situation, it has been repeatedly suggested that digital tools might represent a viable opportunity to improve the efficiency and quality of service delivery, as well as to enhance patient outcomes and experience [14-17].

Previous studies have explored the use of digital solutions in health care settings, such as artificial intelligence (AI)-based interventions and conversational agents. However, these studies have mainly focused on treatment support or remote monitoring [18]. Moreover, there is little evidence of the efficacy of such tools in real-world clinical settings [18,19]. Within the field of mental health care, the use of AI and conversational agents has mainly focused on self-care tools [20], whereas the efficacy of AI in supporting clinicians in their delivery of high-quality care has not been explored. The use of AI is well suited to address the supply-side issues faced by mental health care providers by improving the allocation of staff time to boost service capacity through the support and augmentation of clinicians [21,22]. For example, AI can enable health care professionals to prioritize tasks and streamline processes by automating low-level clinical functions such as adaptive information gathering to inform assessment or treatment sessions conducted by a trained clinician.

Digital innovation to support referral and clinical assessment is earmarked as a key area to increase service capacity within mental health care. One of the main aims of the referral process is to collect information that can be used for clinical assessment to identify symptoms and triage patients into the appropriate treatment pathways. Therefore, the referral process and clinical assessments represent promising targets for automation. These early parts of the care pathway are typically conducted by trained mental health professionals and require considerable time from these overburdened clinical staff. Indeed, studies have found that NHS Talking Therapies (previously Improving Access to Psychological Therapies (IAPT)) services spend up to 25% of their annual budget on clinical assessments [23]. Automation in this area represents a viable opportunity to release clinical time and resources that can be reallocated to other stages of the care pathway.

In addition to service efficiency, other patient benefits can be generated through the implementation of AI-enabled digital solutions. Direct benefits include reduced barriers to entry, such as social stigma and time constraints [24], resulting in a more accessible and patient-focused referral process. In addition, previous research suggests that patients are more likely to report severe symptoms in digital solutions [25], which can lead to more accurate referral information. As a result, clinicians receive a more comprehensive overview of the problems faced by their patients. This presents an opportunity to accelerate clinical assessment, improve pathway allocation, and spend more time during clinical contacts to focus on building a strong relationship with the patient. Indirectly, increased overall efficiency of the service will free up resources that can be reallocated to increase the number of available treatment sessions, which is known to improve clinical outcomes [26].

Therefore, we hypothesize that the use of an AI-enabled referral tool compared with other means of referral will reduce assessment times, reduce wait times for assessment and treatment, reduce dropout rates, reduce changes in treatment allocation, and improve recovery rates. Moreover, we hypothesize that these effects should be largely driven by the collection of clinically relevant information, which can provide valuable context for clinicians at assessment.

Objectives

In this study, we evaluated the impact of an AI self-referral tool, a conversational AI chatbot (Limbic Access [Limbic Limited]), in a real-world scenario. This AI self-referral tool is already implemented as part of routine care across multiple NHS Talking Therapy services in England. We analyzed data from 1 service provider with Talking Therapy services across England. Data were collected from 64,862 patients who were referred for care either via the AI self-referral tool or via alternative methods of referral. We show that the AI solution improves clinical efficiency, reduces wait times and dropout rates, provides more accurate treatment allocation, and increases recovery rates. We further show that frontloading the collection of clinically

relevant information ahead of the clinical assessment is a major driver for these observed improvements. Therefore, our findings provide novel empirical evidence that mental health care can be significantly improved through AI solutions that support trained clinicians in their daily work.

Methods

AI Self-Referral Tool

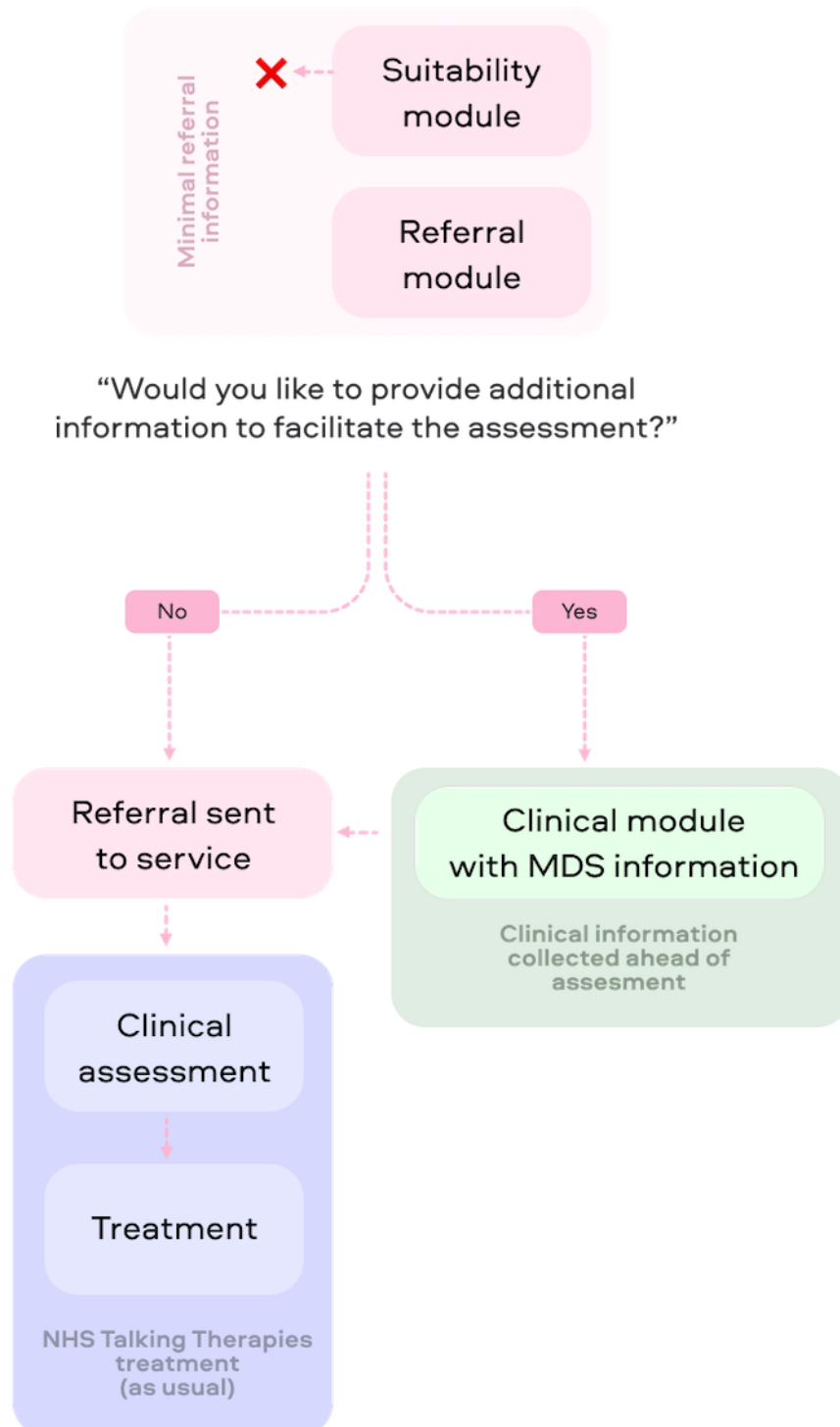
In this study, we evaluated the effects of a novel AI self-referral tool (Limbic Access), which was implemented as part of routine mental health care in several NHS Talking Therapy services. Limbic Access is a commercial product and was developed and commercialized by some of the authors in collaboration with NHS Talking Therapy services. This tool was initially tested in a pilot study with a sample of 7176 patients with 1 NHS Talking Therapy provider. After the successful completion of this pilot study, the tool was rolled out commercially across multiple NHS Talking Therapy providers.

This self-referral tool is a conversational chatbot integrated into the service's website and assists patients in making a referral by collecting the necessary intake information as required by the Talking Therapy program (eg, eligibility criteria, contact details, and demographic information). Furthermore, the chatbot collects additional clinical information about the patient's presenting symptoms, such as the Patient Health Questionnaire-9 (PHQ-9) [27], Generalized Anxiety Disorder Assessment-7

[28], Work and Social Adjustment Scale [29], and a selection of additional screening questions. These routine outcome measures and screening questions are typically not collected at the point of referral in NHS Talking Therapies. All the information collected by the AI self-referral tool is then attached to the referral record within the Talking Therapy service's electronic health record to support clinicians in preparing a high-quality and high-efficiency clinical assessment.

It is important to note that when guiding a patient through referral to Talking Therapies, the AI tool uses a *checkpoint*, where there exists a point at which the patient has provided minimal information required to submit a referral. At this checkpoint, all the required information to submit the patient's referrals to the service was collected. However, patients were then asked whether they would like to provide additional *clinical* information regarding their mental health issues, which was specifically designed to facilitate a clinician-led assessment (Figure 1). This additional information includes free-text input regarding the patient's presenting symptoms as well as standardized, clinically validated routine outcome measures and screening questions. Empirically, most patients choose to provide additional information (approximately 97% of referrals); however, a subset of patients only provided minimally required information at referral (approximately 3% of referrals). This allowed us to implement a quasi-experimental design to test the effects of collecting clinical information on patient treatment outcomes.

Figure 1. Pathway of the AI self-referral tool. The tool is embedded on National Health Service (NHS) Talking Therapies service's web page and pops up when a potential patient navigates to that page. Upon initiating an interaction with the chatbot, the eligibility of the patient is determined in the eligibility module. If ineligible, the patient is signposted out of the service (indicated with a red cross mark). The signposting is based on the same standard characteristics that would be applied in other referral pathways, such as patient's location and age, to ensure that only patients from the service's catchment area and patients who are suited in terms of age will be referred. This ensures that patients do not complete the whole referral process to then be signposted elsewhere later on. Signposting out is unrelated to their mental health symptoms. The eligible patient then continues through the referral module which produces the minimal data set needed to refer the patient to the Talking Therapies service. After the referral module, the patient is asked whether they would like to provide additional information. If they consent, they fill in additional information regarding their mental health issues, which is added to the referral record sent to the Talking Therapies service. If they disagree, their referral is sent directly to the service. MDS: minimum data set.



Clinical Implementation of the AI Self-Referral Tool

To derive maximal clinical value from an AI self-referral tool, appropriate implementation of this tool within a wider service environment is of critical importance. Indeed, the realized benefits of any digital tool rely on how it is used in practice.

Within the evaluated psychotherapy service (Everyturn Mental Health), clinical information collected by the AI self-referral tool was used to triage the severity of patient case presentations (eg, mild, moderate, and severe cases of depression can be differentiated based on the magnitude of the PHQ-9 score). Case presentation, symptom severity, and any associated risk factors are then used by the service to schedule an appropriate duration for a clinician-led assessment (ie, complex or severe cases require longer assessment slots and simpler or mild cases may only require shorter assessment slots). In this way, the NHS Talking Therapy service can use the clinical information to allocate clinical resources in a tailored and efficient manner.

The psychotherapy service additionally enabled a “direct booking” feature within the AI self-referral tool, which provided a means for patients to directly book a preferred time for their clinician-led assessment in the service’s calendar, thus reducing the administrative burden on the service and enabling faster access to a clinical assessment. This might be one mechanism by which this novel referral pathway could reduce wait time for patients.

Finally, all clinical information collected in the AI self-referral tool is programmatically transferred to the service’s chosen patient management system, which can be accessed by the clinician leading the clinical assessment. This provides support to the reviewing clinicians with richer contextual information.

We believe that these implementation decisions for an AI self-referral tool are crucial to consider with respect to the expected effects on service efficiency and quality of care.

Design

Real-world data were collected from patients entering and receiving mental health care treatment through one specific provider of NHS Talking Therapy services (Everyturn Mental Health) between November 2021 and August 2022. The participating mental health services comprised 9 individual Talking Therapy services in different regions throughout England. This allowed us to include data from patients representing diverse geographic and demographic backgrounds (refer to [Multimedia Appendix 1](#) for details on the demographic characteristics of the sample).

In this study, we examined the between- and within-group effects of this AI self-referral solution. In the between-group context, we compared patients who referred themselves to Talking Therapy services through the AI tool with those who were referred through other methods (eg, telephone referrals, referrals via a web form, general practitioner referrals, and referrals via other primary health care services). A comparison of these 2 groups was made possible because of the constant availability of alternative self-referral methods alongside the AI self-referral tool. Overall, these data comprised 64,862 patients, of whom 21,568 (33.25%) patients were referred

through the AI self-referral tool and 43,294 (66.75%) patients were referred through alternative routes.

In the within-group context, we compared users referring through the AI self-referral tool who also completed the full clinical information (clinical information group: 20,860/21,546, 96.82% patients) with those who only completed the minimally required information for a referral (no clinical information group: 686/21,546, 3.18% patients). This allowed for a comparison of the effects of providing clinical information ahead of the assessment to evaluate some of the mechanisms by which the AI self-referral tool achieved its effects. Minimal referral information was defined as patients not completing all relevant clinical information asked for in the self-referral process. It was expected that only a small proportion of patients would not provide complete clinical information, as the AI self-referral tool was designed to increase engagement and ensure that a maximum number of patients complete all relevant information ahead of the clinical assessment.

Ethical Considerations

As determined by the NHS and in accordance with National Institute of Health and Clinical Excellence principles [30], clinical audit studies within the NHS Talking Therapy framework do not require additional patient consent or ethical approval [30]. Moreover, the study team received written confirmation from the Health Research Authority of England that this study constitutes a service evaluation and, therefore, did not require additional ethical approval. When registering to use the AI self-referral tool, patients provided written informed consent as part of a privacy policy agreement, allowing the service to use anonymized patient data for auditing purposes and to support research.

Outcome Measures

The outcome measures reported in this study were assessed routinely during mental health care delivered by NHS Talking Therapy services. Anonymous data were publicly reported on the NHS Digital website [31] for the evaluation of NHS Talking Therapy services performance. Therefore, no additional data beyond routine care data were collected for this study.

Assessment Duration

We evaluated whether the use of the AI self-referral tool improved clinical efficiency by reducing the time required to complete a high-quality clinical assessment. The required length of clinical assessment was measured in minutes.

Wait Time for Clinical Assessment

We evaluated whether the use of the AI self-referral tool reduced the wait time for clinical assessment. The required wait time for clinical assessment was measured in days, from the day of referral to the day of the clinical assessment.

Wait Time for Treatment

We evaluated whether the use of the AI self-referral tool reduced the wait time to the start of treatment. The wait time for treatment was measured in days, from the day of referral to the day of the first treatment session. Only the data of patients who entered treatment were used for this analysis because, for some

patients in the clinical assessment, it might be decided that no treatment is required.

Dropout Rate

We determined whether the use of the AI self-referral tool would reduce the likelihood of patients dropping out of the service at any point during the care pathway. Dropouts were defined as those patients who canceled an appointment and did not rebook a new appointment. The dropout rate was measured as the percentage of patients who dropped out of the treatment.

Change in Allocated Treatment Level

We evaluated whether the use of the AI self-referral tool would enable more accurate clinical assessment. A more accurate clinical assessment would manifest in patients assigned to the appropriate treatment pathway; therefore, the treatment pathway would be less likely to change during treatment. Changes in treatment are known as stepups and stepdowns in NHS Talking Therapies. We measured the accuracy of treatment allocation as the percentage of patients whose treatment was stepped up or down. Only data from patients who received and finished treatment were used for this analysis because the accuracy of treatment allocation can only be assessed after the treatment ends.

Recovery Rate

We evaluated whether the use of the AI self-referral tool would enable a higher rate of recovery in the Talking Therapy service. The recovery of patients is assessed at the end of treatment, and the definition of reliable recovery is systematically used in NHS Talking Therapy services [32]. This was measured by administering an appropriate disorder-specific outcome questionnaire and was defined as a significant reduction in symptom scores (ie, PHQ-9 score: improved by at least 6 points and Generalized Anxiety Disorder Assessment-7 score: improved by at least 4 points) from the beginning to the end of treatment and a score below the clinical cut-off at the end of treatment. We measured the recovery rate as the percentage of patients who achieved reliable recovery. Only data from patients who received and finished treatment were used for this analysis because reliable recovery could only be assessed after completion of the treatment.

Analysis

For the analysis of wait time for treatment, we only analyzed data from patients who had entered treatment. We included patients who had finished their treatment for the changes in treatment allocation and recovery rate analyses.

Because this was not a randomized controlled trial, there may have been differences in the characteristics of the patients referring through the AI tool versus the standard pathway, as well as *within* the AI self-referral tool cohort between patients with clinical information and patients without clinical information. Therefore, we statistically controlled for these potential differences to ensure that our observed results could not be explained by these confounding factors.

The confounding factor of main concern was the severity of patients' mental health symptoms. These data were included for every patient, allowing us to control for this confounding

factor when comparing the AI tool and standard referral pathways. We measured severity as the step of treatment level that patients were assigned to and controlled for severity in any analysis we conducted.

There was only limited information about the group of patients with other referral pathways available to ensure the anonymity of this group. No demographic information or any personally identifiable information was provided for these patients to ensure complete anonymity of data. Therefore, we were unable to control for demographic differences or any other personal information in this data group.

Demographic information was available for patients who were referred through the AI tool. Therefore, for comparison of patients who did and did not provide complete clinical information (all referred via the AI self-referral tool), all analyses controlled for a list of demographic variables (eg, age, gender, ethnicity, disability status, and receiving previous mental health support).

To adequately control for the above-mentioned covariates, we constructed multiple linear regression models for continuous outcome measures and multiple logistic regression models for binary outcome measures. The group was used as a predictor variable (AI vs standard referral comparison: 0=standard referral and 1=AI self-referral; clinical information vs no clinical information comparison: 0=no clinical information and 1=clinical information), and severity was included as a covariate. For the clinical assessment time, wait time to clinical assessment, wait time for treatment, and severity and demographics were the only potentially confounding effects that we controlled for.

Regarding dropout rates, it is possible that increased assessment and wait times could have indirectly led to increased dropouts. Therefore, we controlled for severity and demographics, assessment, and wait time as covariates in the logistic regression model to predict the dropout rates. This analysis will reveal whether the effects on dropout rates are completely explained by the changes in assessment and wait time or whether the use of the AI self-referral tool has an additional and independent effect on dropout rates.

Changes in treatment allocation could potentially be influenced by all the factors mentioned above, including dropout rates. Therefore, we controlled for severity and demographics, dropout rates, assessment, and treatment times in the logistic regression to predict changes in treatment allocation.

Finally, the recovery rate is the last measure of interest, which, in principle, could be influenced by all the factors mentioned above. In particular, changes in treatment allocation (ie, accuracy with which treatment allocation was assigned) could potentially explain why differences in recovery rates were observed. To evaluate whether the effects on the recovery rate could be explained by effects on these other variables or whether it represented independent and additional effects of the AI solution, we included severity and demographics, assessment time, wait time, dropout rates, and changes in treatment allocation as covariates in the logistic regression predicting recovery rates.

Qualitative Analysis on the Reasons to Provide Clinical Information

To investigate the impact of clinical information on relevant outcome measures, we compared patients who provided clinically relevant information to those who did not provide this information.

Because this comparison was a quasi-experimental setup (ie, patients were not randomized into the conditions), we aimed to understand in more detail why patients chose to provide or not provide clinical information.

For this purpose, we analyzed qualitative data from a previous user experience study (unpublished) in which 32 ex-patients tested the AI self-referral tool and answered a subsequent survey on their experience with it. The original focus of this study was to identify potential weaknesses in the design of the AI self-referral tool, thus emphasizing ways to improve the product. This survey included the user experience questionnaire [33] and qualitative feedback questions about their experience. For this purpose, in this study, we focused on the qualitative feedback of the users. We first performed reflexive thematic analysis on feedback entries [34], with 2 of the authors open-coding all the feedback samples. The initial codes were discussed with the larger group of authors, and a consensus was reached on the resulting themes after 2 meetings. The list of resulting themes included comprehension, information about the further steps, ease of user interface use, number of questions, heavy nature of questions, ease of access, and the advantages of the tool's human-free nature. Finally, 2 researchers coded each feedback sample with one of the themes, and the frequency of each category was analyzed. Therefore, we specifically focused on the frequency of the number of questions and the "heavy" nature of questions themes because these were related to the collection of clinically relevant information.

Results

Between-Group Results: Patient Referrals Made via the AI Tool Versus Alternative Routes

We first tested whether the groups of patients were comparable in terms of their severity of mental health conditions. The groups differed in their severity (Mann-Whitney U test; $P<.001$). Patients referred through the AI self-referral tool showed slightly lower severity (mean step of care=1.5) than those referred through other pathways (mean step of care=1.69). Although this was expected based on anecdotal evidence that patients referred through standard pathways show higher severity than patients referred through the AI tool, this finding indicates that it is critical to control for severity in the subsequent analyses.

Assessment Time

A major aspect of an AI self-referral tool is the clinical efficiency generated through this product by reducing the time needed for a clinical assessment. Indeed, in the AI group (mean assessment time=41.6 min), the clinical assessment required, on average, 12.7 minutes less time (Multimedia Appendix 2) compared with the standard referral pathway group (mean assessment time 54.4 minutes). This effect was statistically

significant ($t_{64,861}=-116.57$; $P<.001$) and could not be explained by differences in severity because the effect remained significant after controlling for this factor ($P<.001$). This finding indicates that the use of AI in the self-referral process creates clinical efficiency by reducing clinical assessment times.

Wait Time for Clinical Assessment

Then, we investigated whether the AI self-referral tool affected the time that patients had to wait for their clinical assessment. Indeed, in the AI group, the wait time for a clinical assessment was shorter (mean 15.2 days; Multimedia Appendix 2) than that for the standard referral pathway group (mean 17.4 days). This effect represented an average reduction in wait time of 2.2 days and was statistically significant ($t_{64,861}=-14.66$; $P<.001$). This effect could not be explained by differences in severity because the effect remained significant after controlling for this factor ($P<.001$). This finding indicates that the AI tool reduced the wait times for clinical assessments.

Wait Time to Treatment

Further, we investigated whether the AI self-referral tool affected the time patients had to wait until the first treatment session. In the AI group, the wait time for the first treatment session was shorter (mean 75.6 days; Multimedia Appendix 2) than that for the standard referral pathway group (mean 80.6 days). This effect represented an average reduction in wait time of 5 days and was statistically significant ($t_{33,269}=-7.1$; $P<.001$). This effect could not be explained by differences in severity because the effect remained significant after controlling for this factor ($P<.001$). This finding indicates that the AI tool reduced wait time for accessing mental health treatment.

Dropout Rate

Then, we investigated whether the AI self-referral tool affected the probability of patients dropping out of the treatment. The probability of dropping out of treatment was significantly reduced ($t_{33,269}=9.03$; $P<.001$) from 26.7% probability in the standard referral pathway group to 21.9% probability in the AI tool group (Multimedia Appendix 2). This effect could not be explained by differences in severity or assessment and wait times because the effect remained significant after controlling for this effect ($P<.001$). This finding indicates that the use of the AI tool in the self-referral process reduced the likelihood of patients dropping out of the treatment pathway.

Change in Allocated Treatment Level

Subsequently, we investigated whether the AI self-referral tool affected the accuracy of clinical assessment by investigating the effects on the changes in treatment allocation (ie, the lower rate of change equals improved accuracy of clinical assessment). Changes in treatment allocation were significantly reduced ($t_{20,317}=-8.290$; $P<.001$) from 10.5% of patients receiving a change in treatment in the standard referral pathway group to 5.8% in the AI tool group (Multimedia Appendix 2). This effect could not be explained by differences in severity, dropout rates, or assessment or wait times because the effect remained significant after controlling for these factors ($P<.001$). This finding indicates that the AI self-referral tool improved clinical

assessment accuracy, thus requiring fewer changes in allocation during treatment.

Recovery Rates

Finally, we investigated whether the AI self-referral tool affected the recovery rates of patients. Indeed, in the AI group (recovery rate=58%), the recovery rates were significantly higher ($t_{20,317}=38.7$; $P<.001$; [Multimedia Appendix 2](#)) than those in the standard referral pathway group (recovery rate=27.4%). The effect size is noteworthy as the recovery rate was twice as high in the AI group compared with the standard referral pathway group. This effect could not be explained by differences in severity, dropout rates, assessment and wait times, or by changes in treatment allocation because the effect remained significant after controlling for these factors ($P<.001$). This finding indicates that the use of AI tool in the referral process improved the recovery rates of patients referred through this tool in addition to the other effects presented in this study.

Within-Group Results: Effect of Additional Clinical Information Collected Ahead of Clinician-Led Assessment

Having established the effects of referring through an AI self-referral tool compared with other methods of referral, we investigated more closely the mechanism through which these improvements were achieved. Our initial hypothesis was that the provision of clinically relevant data ahead of the assessment would enable clinicians to better prepare their assessment and create efficiency in their management of the clinical assessment, further enabling them to arrive at accurate clinical conclusions. To test this hypothesis, we investigated only patients referred through the AI self-referral tool, comparing patients who had provided clinical information in their referral to those who provided no clinical information.

First, we ensured that the patient groups did not differ with respect to the most relevant characteristics. Indeed, the groups did not differ with respect to severity (Mann-Whitney U test; $P=.17$), age (Mann-Whitney U test; $P=.42$), gender (Mann-Whitney U test; $P=.44$), ethnicity (Mann-Whitney U test; $P=.39$), disability status (Mann-Whitney U test; $P=.62$), or previous mental health treatment (Mann-Whitney U test; $P=.76$). This finding indicated that the groups were largely comparable. Nevertheless, we included these variables as covariates in the following analyses to ensure that even subtle differences were controlled for.

For the group in which additional clinical information was provided (mean assessment time 40.6 minutes), the clinical assessment required, on average, 12.3 minutes less time compared with the group without clinical information (mean assessment time 52.8 minutes). This effect was statistically significant ($t_{21,545}=-16.16$; $P<.001$; [Multimedia Appendix 3](#)), and this could not be explained by differences in severity or demographics because the effect remained significant after controlling for these factors ($P<.001$).

Furthermore, in the group of patients with clinical information, the wait time for clinical assessment was shorter (mean 15 days)

than that in the group without clinical information (mean 20.2 days).

This effect represented an average reduction of wait time of 5.2 days and was statistically significant ($t_{21,545}=-9.7$; $P<.001$; [Multimedia Appendix 3](#)) and could not be explained by differences in severity or demographics because the effect remained significant after controlling for these factors ($P<.001$).

Finally, in the group with clinical information (recovery rate=58.7%), the recovery rates were significantly higher ($t_{5990}=2.3$; $P=.02$; [Multimedia Appendix 3](#)) than in the group without clinical information (recovery rate=46.9%). This effect could not be explained by differences in severity, demographics, dropout rates, assessment and wait times, or by changes in treatment allocation because the effect remained significant after controlling for these factors ($P=.03$).

Notably, there were also some effects that seemed not to be driven by the clinical information provided ahead of time. There were no significant differences between patients with and without clinical information regarding dropout rates ($P=.26$), wait time for treatment ($P=.51$), and allocation to the accurate treatment level ($P=.86$). This finding suggests that the use of an AI self-referral solution improves access and treatment, with some of its effects being specific to the provision of high-quality symptom data to a clinician.

Qualitative Analysis of the Reasons to Provide Clinical Information

We compared patients who provided all clinical information with those who did not provide this information to evaluate the impact of this clinical information on treatment outcomes.

However, because this study was a quasi-experimental setup, we aimed to understand why the patients chose to provide clinical information or not. To do so, we analyzed qualitative user research with 32 ex-patients to understand their experience with the AI self-referral tool using reflexive thematic analysis techniques. In this analysis, we focused on topics related to clinical information, with 2 relevant recurring topics identified. First, 38% (12/32) of the patients reported that the number of questions was perceived as long and potentially overwhelming. Second, 25% (8/32) of the patients reported that the nature of the clinical questions was emotionally difficult and could feel too heavy to complete.

This finding indicates that one of the main reasons for not providing clinical information might be time constraints and the feeling of being overwhelmed by providing detailed clinical information during referral.

However, the participants were ex-patients who were not seeking to refer themselves to treatment at the point of the study, which might make the collection of this information less directly relevant to them. Moreover, it is important to note that this study focused on the potential weaknesses of tool design. These results can be complemented by an analysis of 42,332 patients providing qualitative feedback after using the AI-referral tool in a real-world setting reported by Habicht et al [35]. In that analysis, 89% of the patients reported positive feedback on tool use, whereas only 7% gave neutral feedback, and 4% gave

negative feedback. Notably, none of the negative feedback categories included complaints regarding the length or emotional content of the questions. This finding indicates that problems with the number of questions and their emotional content are rare in a real-world setting and might be more apparent when participants are pressed to suggest potential improvements. This finding is in line with the small number of patients not providing clinical information in our study.

Discussion

Principal Findings

In this study, we investigated the effects of implementing an AI self-referral tool in referral and assessment processes for mental health care. To this end, we compared patients referred through this AI tool against those referred through other means of referral within the same NHS Talking Therapy services and in a comparable time frame. In doing so, we demonstrated the improved service efficiency and clinical efficacy associated with this novel tool. Moreover, we investigated the mechanism through which these improvements were achieved and found that the provision of clinical information ahead of the mental health assessment was critical for many of the observed effects.

We found that patients accessing care through the AI tool showed reduced time required to complete their clinician-led assessment, reduced wait times for the assessment and treatment sessions, reduced dropout rates, improved accuracy of treatment allocation, and improved recovery rates. Moreover, we showed that the reduced assessment times, reduced wait times for assessment, and increased recovery rates were largely driven by the additional clinically relevant information collected from patients during their referral via the AI tool. Although our chatbot was friendly but not optimized to express compassion, the increase in efficiency can be seen as compassion for patients' time and resources [36]. Although the effect of clinical information is more straightforward to explain for assessment time and recovery rates, we also observed an effect on wait times, which might appear less intuitive. The likely reason for this effect is a direct booking feature in the AI-referral tool, in which patients can immediately book an appointment in the services' patient management system. However, this feature is only available once patients have provided all clinical information (ie, at the end of the referral process) to allow simple triage and assignment to the appropriate type of assessment (eg, question 9 of the PHQ-9 is required to assess suicidal ideation and thus associated risk). Therefore, this feature is not available for patients who did not provide clinical information.

It is important to note that we conducted multiple control analyses to rule out confounding factors and to establish the independence of these observed effects. Importantly, the severity of cases could not explain the differences between people referred through the AI tool compared with standard referrals. This finding is particularly important because any difference in recovery rates could be expected to be driven by symptom severity; therefore, we have ensured that the improvement seen by the AI self-referral tool cannot be explained by symptom severity. Other potentially confounding factors (eg, users of a

new AI solution may have been more motivated to engage in therapy than patients referred by their general practitioner) are beyond the scope of our analyses and cannot be conclusively ruled out. Nevertheless, other studies evaluating the AI self-referral tool (Limbic Access) have also shown overall positive effects on provider level [37], that is, showing that NHS Talking Therapy providers using this tool showed overall increased recovery rates compared with matched Talking Therapy providers not using the tool. If a selection bias was the explanation for the observed effects, this would suggest no overall improvement in treatment outcomes for providers using the tool. Thus, findings from this related study [37] make a selection bias highly unlikely as an explanation for the observed benefits of the AI tool.

A randomized controlled trial is the gold standard for further confirming the observed effects of this study. However, randomized controlled trials have their shortcomings because they are costly to run and, therefore, limit the available sample size. We chose our experimental design to allow us to investigate an unprecedentedly large sample, yielding high statistical power and excellent ecological validity for our findings. Moreover, because our comparison is based on referrals within the same NHS Talking Therapy service, representing multiple geographies, our findings are unlikely to be driven by differences in demographic variables or general factors, such as geography, and should, therefore, be transferred to other Talking Therapy services.

In addition, we carefully tested whether all the observed effects were independent of each other. All reported effects remained significant when controlling for mutual influences, indicating that using the AI tool in the referral process has beneficial effects on all the variables reported in this study.

We investigated the mechanisms by which the AI self-referral tool improves clinical efficiency. We demonstrated that the provision of clinical information in referrals may be an important component of the observed effects. More specifically, we found that patients who provided clinical information during their referral had reduced assessment times, reduced wait times for assessment, and increased recovery rates. This finding indicates that the provision of clinical information ahead of clinical assessment could be a critical ingredient through which the AI tool achieved its effect on the tested outcome measures. This finding was hypothesized and showed that an increased amount of relevant information for the preparation of the clinical assessment has beneficial effects on patients and IAPT services.

In contrast, it is notable that not all effects observed for the AI solution (compared with other means of referrals) appeared to be driven by the provision of clinical information ahead of the clinical assessment. This might be expected for some of these effects. For instance, the reduction in dropout rates might be driven more by an overall positive experience that patients have when engaging with a friendly chatbot for submitting a referral, independent of the clinical information provided. Similarly, reductions in wait time for treatment might be driven more by the general administrative burden and overall resource availability rather than the specific clinical information provided in the referral.

However, it is notable that the provision of clinical information did not seem to have a significant effect on the accuracy of treatment allocation. This effect would have been expected to benefit from clinical information ahead of the clinical assessment. Nevertheless, there are 2 points to be considered with respect to this finding. First, there were a small number of patients (153/21,568, 0.71%) who did not provide clinical information and finished their treatment in this study, which dramatically reduced the power of the analysis compared with the analysis looking at the general effects of the AI solution compared with standard pathway referrals. Therefore, the nonsignificant results could be partly explained by the noise in a small sample. Second, it is notable that although the clinical information provided in this version of the AI tool is useful for many aspects of the clinical assessment process, it is fairly generic, mainly covering information about depression, generalized anxiety, and functional impairment. Although this information is useful in allocating accurate resources in the assessment and in prioritizing severe cases, it only provides limited information about the more specific symptoms that the patient experiences. This is especially true when the patient is experiencing mental health problems that do not represent depression or generalized anxiety. Therefore, the provision of more tailored and specific information at the point of referral would likely yield better results and support improvements regarding the allocation of treatment pathways.

Limitations

Although this study aimed to maximize ecological validity and power using a large sample real-world data set, this decision has some limitations. As discussed above, this study was an observational study using a quasi-experimental setup. This means that the participants were not randomly allocated to each study arm (ie, type of referral). Although a multitude of control analyses have been conducted to ensure that the observed effects were not confounded by different characteristics of the patients (eg, case severity), it is not possible to measure and control for all potential confounding factors. Therefore, there remains the possibility of confounding factors between the study arms.

Moreover, we investigated the effects of clinical information and the observed benefits of the AI-enabled referral tool. Further, this was investigated using a quasi-experimental setup, which could have led to some form of confounds, even though careful statistical control of different characteristics has been conducted. It is noteworthy that in a separate usability study,

patients reported that the self-referral process can be long and emotionally difficult, indicating that patients not providing clinical information could have done so because of time constraints or emotional burden. It is possible that these characteristics (eg, reduced time capacity or difficulties in facing emotional topics) could interact with treatment success and could influence the observed effects, such as improved recovery rates. Although we controlled for many confounding factors, it was not possible to further control for these potential effects and to conclusively rule out this possibility.

Conclusions

This study represents, to the best of our knowledge, first evidence of the real-world impact of an AI-enabled self-referral tool in mental healthcare. The study was conducted with a large sample of patients in a mental health care setting, yielding a high ecological validity of the reported findings.

Notably, the results indicated a strong positive real-world impact of this novel AI tool (Limbic Access) on clinical efficacy and efficiency.

The setup for this study was quasi-experimental, so that not all confounding factors could be controlled completely. However, we assessed and controlled for the most relevant factors that could have differed between the groups of comparison. Notably, none of these factors could explain the observed effects, and all the effects remained significant after controlling for these factors.

It is critical to note that we provided converging evidence from multiple sources of data and different analyses. We conducted multiple control analyses to derive the most reliable and robust conclusions. Nevertheless, as none of the analyses included a randomized controlled trial, the possibility of confounding factors remained even though we controlled for most factors. Notwithstanding, the different analyses had different strengths and weaknesses, and no confounding factors could explain all the observed results.

Therefore, the results highlight the specific, beneficial role that well-designed AI solutions can play in augmenting the work of human clinicians by supporting elements of clinical work and through this, freeing up time for clinicians. This means that AI solutions can enable mental health care providers to deal with increased demand, even within a challenging funding environment that precludes increases in staffing levels.

Acknowledgments

The authors wish to thank Emili Ivanova for her contribution to the figures and illustrations. The authors wish to thank the Everyturn Mental Health Talking Therapy team for their support in data acquisition for this study.

Data and Code Availability

All analyses were based on sensitive clinical data, which we are not permitted to share with third parties. Summary statistics, as displayed in figures and text, are available upon request from the corresponding author.

Conflicts of Interest

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.

Multimedia Appendix 1

Demographic characteristics of patients using the artificial intelligence (AI)-enabled referral tool and those who did not use the AI tool (ie, referred through other means). Although for patients using the AI tool, data were collected during the self-referral process, there were no individual-level demographics available for patients who did not use the AI tool. Group-level demographics were acquired using data from the National Health Service Digital database.

[[PDF File \(Adobe PDF File\), 36 KB - ai_v2i1e44358_app1.pdf](#)]

Multimedia Appendix 2

Comparison of treatment outcomes between referrals through the artificial intelligence (AI) self-referral tool versus standard referrals: (A) assessment time (in min), (B) wait time from referral to assessment (in days), (C) wait time from referral to first treatment session (in days), (D) dropout rates from treatment, (E) change in treatment level (measured as stepups and stepdowns in treatment level), and (F) recovery rate (ie, reliable recovery). Error bars indicate SEs. Because of the large sample size, some SEs are very small and thus hard to see. *** $P < .001$.

[[PNG File , 170 KB - ai_v2i1e44358_app2.png](#)]

Multimedia Appendix 3

Comparison of treatment outcomes between artificial intelligence tool referrals with and without clinical information: (A) assessment time (in min), (B) wait time from referral to assessment (in days), and (C) recovery rate (ie, reliable recovery). Error bars indicate SEs. Because of the large sample size, some SEs are very small and thus hard to see. *** $P < .001$ and * $P < .05$.

[[PNG File , 116 KB - ai_v2i1e44358_app3.png](#)]

References

1. Nochaiwong S, Ruengorn C, Thavorn K, Hutton B, Awiphan R, Phosuya C, et al. Global prevalence of mental health issues among the general population during the coronavirus disease-2019 pandemic: a systematic review and meta-analysis. *Sci Rep* 2021 May 13;11(1):10173 [FREE Full text] [doi: [10.1038/s41598-021-89700-8](https://doi.org/10.1038/s41598-021-89700-8)] [Medline: [33986414](https://pubmed.ncbi.nlm.nih.gov/33986414/)]
2. Horackova K, Kopecek M, Machů V, Kagstrom A, Aarsland D, Motlova LB, et al. Prevalence of late-life depression and gap in mental health service use across European regions. *Eur Psychiatry* 2019 Apr 15;57:19-25. [doi: [10.1016/j.eurpsy.2018.12.002](https://doi.org/10.1016/j.eurpsy.2018.12.002)] [Medline: [30658276](https://pubmed.ncbi.nlm.nih.gov/30658276/)]
3. Arias De La Torre JA, Vilagut G, Ronaldson A, Serrano-Blanco A, Valderas J, Martín V, et al. Prevalence of depression in Europe using two different PHQ-8 scoring methods. *Eur Psychiatry* 2022 Sep 01;65(S1):S299 [FREE Full text] [doi: [10.1192/j.eurpsy.2022.763](https://doi.org/10.1192/j.eurpsy.2022.763)]
4. Busetta G, Campolo MG, Fiorillo F, Pagani L, Panarello D, Augello V. Effects of COVID-19 lockdown on university students' anxiety disorder in Italy. *Genus* 2021 Oct 09;77(1):25 [FREE Full text] [doi: [10.1186/s41118-021-00135-5](https://doi.org/10.1186/s41118-021-00135-5)] [Medline: [34658399](https://pubmed.ncbi.nlm.nih.gov/34658399/)]
5. Murch BJ, Cooper JA, Hodgett TJ, Gara EL, Walker JS, Wood RM. Modelling the effect of first-wave COVID-19 on mental health services. *Oper Res Health Care* 2021 Sep;30:100311 [FREE Full text] [doi: [10.1016/j.orhc.2021.100311](https://doi.org/10.1016/j.orhc.2021.100311)] [Medline: [36466119](https://pubmed.ncbi.nlm.nih.gov/36466119/)]
6. Ornell F, Borelli WV, Benzano D, Schuch JB, Moura HF, Sordi AO, et al. The next pandemic: impact of COVID-19 in mental healthcare assistance in a nationwide epidemiological study. *Lancet Reg Health Am* 2021 Dec 04;4:100061 [FREE Full text] [doi: [10.1016/j.lana.2021.100061](https://doi.org/10.1016/j.lana.2021.100061)] [Medline: [34518824](https://pubmed.ncbi.nlm.nih.gov/34518824/)]
7. Marques L, Bartuska AD, Cohen JN, Youn SJ. Three steps to flatten the mental health need curve amid the COVID-19 pandemic. *Depress Anxiety* 2020 May;37(5):405-406 [FREE Full text] [doi: [10.1002/da.23031](https://doi.org/10.1002/da.23031)] [Medline: [32429005](https://pubmed.ncbi.nlm.nih.gov/32429005/)]
8. Loosen AM, Skvortsova V, Hauser TU. Obsessive-compulsive symptoms and information seeking during the COVID-19 pandemic. *Transl Psychiatry* 2021 May 21;11(1):309 [FREE Full text] [doi: [10.1038/s41398-021-01410-x](https://doi.org/10.1038/s41398-021-01410-x)] [Medline: [34021112](https://pubmed.ncbi.nlm.nih.gov/34021112/)]
9. Thome J, Deloyer J, Coogan AN, Bailey-Rodriguez D, da Cruz E Silva OA, Faltraco F, et al. The impact of the early phase of the COVID-19 pandemic on mental-health services in Europe. *World J Biol Psychiatry* 2021 Sep;22(7):516-525. [doi: [10.1080/15622975.2020.1844290](https://doi.org/10.1080/15622975.2020.1844290)] [Medline: [33143529](https://pubmed.ncbi.nlm.nih.gov/33143529/)]
10. Scott MJ. Improving access to psychological therapies (IAPT) - the need for radical reform. *J Health Psychol* 2018 Aug;23(9):1136-1147 [FREE Full text] [doi: [10.1177/1359105318755264](https://doi.org/10.1177/1359105318755264)] [Medline: [29390891](https://pubmed.ncbi.nlm.nih.gov/29390891/)]
11. Psychological therapies, annual report on the use of IAPT services, 2021-22. NHS Digital. 2022. URL: <https://digital.nhs.uk/data-and-information/publications/statistical/psychological-therapies-annual-reports-on-the-use-of-iapt-services/annual-report-2021-22> [accessed 2023-11-06]

12. Larsson P, Lloyd R, Taberham E, Rosairo M. An observational study on IAPT waiting times before, during and after the COVID-19 pandemic using descriptive time-series data. *Ment Health Rev J* 2022 Jul 27;27(4):455-471 [FREE Full text] [doi: [10.1108/mhrj-04-2022-0023](https://doi.org/10.1108/mhrj-04-2022-0023)]
13. Adams R, Ryan T, Wood E. Understanding the factors that affect retention within the mental health nursing workforce: a systematic review and thematic synthesis. *Int J Ment Health Nurs* 2021 Dec;30(6):1476-1497 [FREE Full text] [doi: [10.1111/inm.12904](https://doi.org/10.1111/inm.12904)] [Medline: [34184394](https://pubmed.ncbi.nlm.nih.gov/34184394/)]
14. Jayaraajan K, Sivananthan A, Koomson A, Ahmad A, Haque M, Hussain M. The use of digital solutions in alleviating the burden of IAPT's waiting times. *Int J Risk Saf Med* 2022;33(S1):S103-S110 [FREE Full text] [doi: [10.3233/JRS-227033](https://doi.org/10.3233/JRS-227033)] [Medline: [35912756](https://pubmed.ncbi.nlm.nih.gov/35912756/)]
15. Rudd B, Beidas R. Digital mental health: the answer to the global mental health crisis? *JMIR Ment Health* 2020 Jun 02;7(6):e18472 [FREE Full text] [doi: [10.2196/18472](https://doi.org/10.2196/18472)] [Medline: [32484445](https://pubmed.ncbi.nlm.nih.gov/32484445/)]
16. Koutsouleris N, Hauser TU, Skvortsova V, De Choudhury M. From promise to practice: towards the realisation of AI-informed mental health care. *The Lancet Digital Health* 2022 Nov;4(11):e829-e840. [doi: [10.1016/s2589-7500\(22\)00153-4](https://doi.org/10.1016/s2589-7500(22)00153-4)]
17. Hauser TU, Skvortsova V, De Choudhury M, Koutsouleris N. The promise of a model-based psychiatry: building computational models of mental ill health. *Lancet Digit Health* 2022 Nov;4(11):e816-e828 [FREE Full text] [doi: [10.1016/S2589-7500\(22\)00152-2](https://doi.org/10.1016/S2589-7500(22)00152-2)] [Medline: [36229345](https://pubmed.ncbi.nlm.nih.gov/36229345/)]
18. Tudor Car L, Dhinakaran DA, Kyaw BM, Kowatsch T, Joty S, Theng Y, et al. Conversational agents in health care: scoping review and conceptual analysis. *J Med Internet Res* 2020 Aug 07;22(8):e17158 [FREE Full text] [doi: [10.2196/17158](https://doi.org/10.2196/17158)] [Medline: [32763886](https://pubmed.ncbi.nlm.nih.gov/32763886/)]
19. 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 Sep 01;25(9):1248-1258 [FREE Full text] [doi: [10.1093/jamia/ocy072](https://doi.org/10.1093/jamia/ocy072)] [Medline: [30010941](https://pubmed.ncbi.nlm.nih.gov/30010941/)]
20. Pham KT, Nabizadeh A, Selek S. Artificial intelligence and chatbots in psychiatry. *Psychiatr Q* 2022 Mar;93(1):249-253 [FREE Full text] [doi: [10.1007/s11126-022-09973-8](https://doi.org/10.1007/s11126-022-09973-8)] [Medline: [35212940](https://pubmed.ncbi.nlm.nih.gov/35212940/)]
21. D'Alfonso S. AI in mental health. *Curr Opin Psychol* 2020 Dec;36:112-117. [doi: [10.1016/j.copsyc.2020.04.005](https://doi.org/10.1016/j.copsyc.2020.04.005)] [Medline: [32604065](https://pubmed.ncbi.nlm.nih.gov/32604065/)]
22. Ćosić K, Popović S, Šarlija M, Kesedžić I. Impact of human disasters and COVID-19 pandemic on mental health: potential of digital psychiatry. *Psychiatr Danub* 2020;32(1):25-31 [FREE Full text] [doi: [10.24869/psyd.2020.25](https://doi.org/10.24869/psyd.2020.25)] [Medline: [32303026](https://pubmed.ncbi.nlm.nih.gov/32303026/)]
23. Scott M. The cost of iapt is at least five times greater than claimed. *CBT watch*. 2018. URL: <http://www.cbtwatch.com/the-cost-of-iapt-is-at-least-five-times-greater-than-claimed/> [accessed 2023-11-06]
24. Lattie EG, Stiles-Shields C, Graham AK. An overview of and recommendations for more accessible digital mental health services. *Nat Rev Psychol* 2022 Jan 26;1(2):87-100 [FREE Full text] [doi: [10.1038/s44159-021-00003-1](https://doi.org/10.1038/s44159-021-00003-1)]
25. Torous J, Staples P, Shanahan M, Lin C, Peck P, Keshavan M, et al. Utilizing a personal smartphone custom app to assess the Patient Health Questionnaire-9 (PHQ-9) depressive symptoms in patients with major depressive disorder. *JMIR Ment Health* 2015 Mar 24;2(1):e8 [FREE Full text] [doi: [10.2196/mental.3889](https://doi.org/10.2196/mental.3889)] [Medline: [26543914](https://pubmed.ncbi.nlm.nih.gov/26543914/)]
26. Gyani A, Shafran R, Layard R, Clark DM. Enhancing recovery rates: lessons from year one of IAPT. *Behav Res Ther* 2013 Sep;51(9):597-606 [FREE Full text] [doi: [10.1016/j.brat.2013.06.004](https://doi.org/10.1016/j.brat.2013.06.004)] [Medline: [23872702](https://pubmed.ncbi.nlm.nih.gov/23872702/)]
27. Kroenke K, Spitzer RL, Williams JB. The PHQ-9: validity of a brief depression severity measure. *J Gen Intern Med* 2001 Sep;16(9):606-613 [FREE Full text] [doi: [10.1046/j.1525-1497.2001.016009606.x](https://doi.org/10.1046/j.1525-1497.2001.016009606.x)] [Medline: [11556941](https://pubmed.ncbi.nlm.nih.gov/11556941/)]
28. Spitzer RL, Kroenke K, Williams JB, Löwe B. A brief measure for assessing generalized anxiety disorder: the GAD-7. *Arch Intern Med* 2006 May 22;166(10):1092-1097. [doi: [10.1001/archinte.166.10.1092](https://doi.org/10.1001/archinte.166.10.1092)] [Medline: [16717171](https://pubmed.ncbi.nlm.nih.gov/16717171/)]
29. Mundt JC, Marks IM, Shear MK, Greist JM. The Work and Social Adjustment Scale: a simple measure of impairment in functioning. *Br J Psychiatry* 2002 May 02;180(5):461-464. [doi: [10.1192/bjp.180.5.461](https://doi.org/10.1192/bjp.180.5.461)] [Medline: [11983645](https://pubmed.ncbi.nlm.nih.gov/11983645/)]
30. Rawlins M. Principles for Best Practice in Clinical Audit. Oxford, UK: Radcliffe Publishing; 2002.
31. Psychological therapies, annual reports on the use of IAPT services. NHS Digital. URL: <https://digital.nhs.uk/data-and-information/publications/statistical/psychological-therapies-annual-reports-on-the-use-of-iapt-services> [accessed 2023-11-06]
32. Jacobson N, Truax P. Clinical significance: a statistical approach to defining meaningful change in psychotherapy research. *J Consult Clin Psychol* 1991 Feb;59(1):12-19. [doi: [10.1037//0022-006x.59.1.12](https://doi.org/10.1037//0022-006x.59.1.12)] [Medline: [2002127](https://pubmed.ncbi.nlm.nih.gov/2002127/)]
33. Schrepp M, Hinderks A, Thomaschewski J. Construction of a Benchmark for the User Experience Questionnaire (UEQ). *Int J Interact Multimed Artif Intell* 2017;4(4):40 [FREE Full text] [doi: [10.9781/ijimai.2017.445](https://doi.org/10.9781/ijimai.2017.445)]
34. Braun V, Clarke V. Reflecting on reflexive thematic analysis. *Qual Res Sport Exerc Health* 2019 Jun 13;11(4):589-597 [FREE Full text] [doi: [10.1080/2159676x.2019.1628806](https://doi.org/10.1080/2159676x.2019.1628806)]
35. Habicht J, Viswanathan S, Carrington B, Hauser T, Harper R, Rollwage M. Closing the accessibility gap to mental health treatment with a conversational AI-enabled self-referral tool. medRxiv. Preprint posted online May 1, 2023. [FREE Full text] [doi: [10.1101/2023.04.29.23289204](https://doi.org/10.1101/2023.04.29.23289204)]
36. Morrow E, Zidaru T, Ross F, Mason C, Patel KD, Ream M, et al. Artificial intelligence technologies and compassion in healthcare: a systematic scoping review. *Front Psychol* 2022 Jan 17;13:971044 [FREE Full text] [doi: [10.3389/fpsyg.2022.971044](https://doi.org/10.3389/fpsyg.2022.971044)] [Medline: [36733854](https://pubmed.ncbi.nlm.nih.gov/36733854/)]

37. Rollwage M, Juchems K, Habicht J, Carrington B, Hauser T, Harper R. Conversational AI facilitates mental health assessments and is associated with improved recovery rates. medRxiv. Preprint posted online May 1, 2023. [[FREE Full text](#)] [doi: [10.1101/2022.11.03.22281887](https://doi.org/10.1101/2022.11.03.22281887)]

Abbreviations

AI: artificial intelligence

IAPT: Improving Access to Psychological Therapies

NHS: National Health Service

PHQ-9: Patient Health Questionnaire-9

Edited by K El Emam, B Malin; submitted 16.01.23; peer-reviewed by M Mulvenna, L Sikstrom; comments to author 15.04.23; revised version received 31.05.23; accepted 20.10.23; published 13.12.23.

Please cite as:

Rollwage M, Habicht J, Juechems K, Carrington B, Viswanathan S, Stylianou M, Hauser TU, Harper R

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

JMIR AI 2023;2:e44358

URL: <https://ai.jmir.org/2023/1/e44358>

doi: [10.2196/44358](https://doi.org/10.2196/44358)

PMID:

©Max Rollwage, Johanna Habicht, Keno Juechems, Ben Carrington, Sruthi Viswanathan, Mona Stylianou, Tobias U Hauser, Ross Harper. Originally published in JMIR AI (<https://ai.jmir.org>), 13.12.2023. This is an open-access article distributed under the terms of the Creative Commons Attribution License (<https://creativecommons.org/licenses/by/4.0/>), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR AI, is properly cited. The complete bibliographic information, a link to the original publication on <https://www.ai.jmir.org/>, as well as this copyright and license information must be included.

Original Paper

The Evolution of Artificial Intelligence in Biomedicine: Bibliometric Analysis

Jiasheng Gu^{1*}, MS; Chongyang Gao², MS; Lili Wang^{3*}, PhD

¹Department of Computer Science, University of Southern California, Los Angeles, CA, United States

²Department of Computer Science, Northwestern University, Evanston, IL, United States

³Department of Computer Science, Dartmouth College, Hanover, NH, United States

*these authors contributed equally

Corresponding Author:

Lili Wang, PhD

Department of Computer Science

Dartmouth College

15 Thayer Drive

Hanover, NH, 03755

United States

Phone: 1 516 888 6691

Email: lili.wang.gr@dartmouth.edu

Abstract

Background: The utilization of artificial intelligence (AI) technologies in the biomedical field has attracted increasing attention in recent decades. Studying how past AI technologies have found their way into medicine over time can help to predict which current (and future) AI technologies have the potential to be utilized in medicine in the coming years, thereby providing a helpful reference for future research directions.

Objective: The aim of this study was to predict the future trend of AI technologies used in different biomedical domains based on past trends of related technologies and biomedical domains.

Methods: We collected a large corpus of articles from the PubMed database pertaining to the intersection of AI and biomedicine. Initially, we attempted to use regression on the extracted keywords alone; however, we found that this approach did not provide sufficient information. Therefore, we propose a method called “background-enhanced prediction” to expand the knowledge utilized by the regression algorithm by incorporating both the keywords and their surrounding context. This method of data construction resulted in improved performance across the six regression models evaluated. Our findings were confirmed through experiments on recurrent prediction and forecasting.

Results: In our analysis using background information for prediction, we found that a window size of 3 yielded the best results, outperforming the use of keywords alone. Furthermore, utilizing data only prior to 2017, our regression projections for the period of 2017-2021 exhibited a high coefficient of determination (R^2), which reached up to 0.78, demonstrating the effectiveness of our method in predicting long-term trends. Based on the prediction, studies related to proteins and tumors will be pushed out of the top 20 and become replaced by early diagnostics, tomography, and other detection technologies. These are certain areas that are well-suited to incorporate AI technology. Deep learning, machine learning, and neural networks continue to be the dominant AI technologies in biomedical applications. Generative adversarial networks represent an emerging technology with a strong growth trend.

Conclusions: In this study, we explored AI trends in the biomedical field and developed a predictive model to forecast future trends. Our findings were confirmed through experiments on current trends.

(JMIR AI 2023;2:e45770) doi:[10.2196/45770](https://doi.org/10.2196/45770)

KEYWORDS

bibliometrics; trend forecasting; AI in medicine; Word2Vec; regression models; agglomerative clustering; usage; artificial intelligence; utilization; biomedical; effectiveness; AI trends; predictive model; development

Introduction

Artificial Intelligence in Biomedicine

Medicine has long been recognized as a prime area for applying artificial intelligence (AI) [1], with biomedicine being a vibrant and promising field. Advances in technology and science have led to the use of various methods to obtain biomedical data, such as clinical analyses, biological parameters, and medical imaging. However, the diversity and complexity of these data, along with the need for more information on certain atypical diseases result in unbalanced and nonsmooth biomedical data. In this scenario, machine learning can improve medical big data analysis, reduce the risk of medical errors, and generate a more unified diagnostic and prognostic protocol.

Recent AI research has leveraged machine learning methods to identify patterns and complex interactions from data, which require large amounts of data as support. Artificial neural networks and deep learning are currently among the most popular machine learning technologies. These methods are used in biomedicine across all medical dimensions, from genomic applications such as gene expression to public health care management such as for predicting population information or infectious disease outbreaks [2]. AI has also significantly impacted biomedical processors such as electrocardiogram, electroencephalogram, and electromyography classification processors and hearing aid processors [3].

AI is increasingly being utilized in a variety of applications in the biomedical field. Notable examples include IBM Watson-Oncology, which selects drugs for cancer treatment with equal or superior efficiency compared to human experts; Microsoft's Hanover project at Oregon, which personalizes cancer treatment plans through analysis of medical research; and the UK National Health Service utilizing Google's DeepMind platform to detect health risks by analyzing mobile app data and medical images from patients. Additionally, algorithms developed at Stanford University have been shown to detect pneumonia more accurately than human radiologists; in the diabetic retinopathy challenge, the computer was as effective as an ophthalmologist in making referral decisions [4]. Therefore, it is essential to analyze the trends in the integration of these AI-related technologies with the biomedical field to understand which technologies have played an important role in the past, predict the current and emerging technologies that are more likely to be important in the future, and determine which original technologies are regaining importance in a particular biomedical field.

Language models offer an effective means to analyze texts and have become the basis for many applications, including machine translation and text classification. In all text-related fields, language models can bring new improvements and opportunities to a greater or lesser extent and assist in literature research.

Co-word Analysis

Recently, increased attention has been paid to the management of references and expansion of the research scope. Bibliometric analysis summarizes the structure of a field by analyzing the social and structural relationships between different research

components such as authors, countries, institutions, and topics. Additionally, bibliometric analysis significantly impacts reorienting research and identifying popular issues. Thus, bibliometric analysis enables discovery of how research in a given field is distributed and changing. The data collected and the conclusions drawn from a bibliometric analysis can be used to track popular topics, predict promising technologies, and assist scientists in redirecting their research. There has been substantial research and application of bibliometric analysis in academia and industry, and extracting keywords to analyze texts is a very common strategy in such studies. Although it is intuitive to use the whole text as an object of analysis, this requires extensive computational resources. Moreover, many texts are not of high quality, some of them are repetitive or have no actual content, and a lot of noise can make the model learn the wrong knowledge. Therefore, keyword-focused analysis is often a better choice. Co-word analysis is one such technique that focuses on keywords and analyzes the content itself [5]. This analysis aims to uncover the intrinsic connections of articles and discover trends within them with applications in many fields, including medicine and business.

Co-word analysis was first proposed by French bibliometricians in the late 1970s [6] as a technique for studying keywords in the content of publications. Words in the co-word analysis are typically derived from the article title, abstract, and full text. These words may be specifically extracted from certain parts of each component, depending on the goal of the analysis. Co-word analysis assumes that words that frequently occur together have thematic relationships with each other. Based on this assumption, co-word analysis can be used to predict future research in a field. Analysis of the keywords of published articles in a given field has the potential to predict keywords for future research in the field, which in turn portrays the future of the research field accordingly. Co-word analysis uses several methods based on covariate matrices, such as factor, cluster, multivariate, and social network analyses. These methods help researchers to obtain an overview of a field. Thus, co-word analysis is a method to analyze papers in a field and make valid judgments.

Text Similarity

Text similarity measurement is fundamental to natural language processing tasks and is essential in information retrieval, question answering, machine translation, and dialogue systems, among other applications. In recent years, various techniques for measuring semantic similarity have been proposed. Text similarity techniques can be divided into two main categories: text distance and text representation [7].

Text distance describes the semantic similarity of two text words from the perspective of distance. Length-based and distribution-based distance are the two main types of text distance. Traditionally, text similarity is evaluated by measuring the length distance, which uses the numerical properties of the text to calculate the text vector distance length, such as the Euclidean distance, cosine distance, or Manhattan distance [8]. However, the text similarity should not be symmetric and the length distance does not consider the statistical characteristics of the data. The distribution distance determines the similarity

between documents based on the similarity of their distribution, such as Jensen-Shannon divergence [9], Kullback-Leibler divergence [10], and Wasserstein distance [11], among others.

Text representation methods convert text to a numerical feature vector. These methods are mainly divided into a string-based method, corpus-based method, semantic text matching, and graph structure-based method. String-based methods operate on string sequences and character compositions to measure the similarity or dissimilarity (distance) between two text strings for approximate string matching or comparison. The advantage of such methods is that they are simple to compute. Representative string-based methods include longest common subsequence [12], Edit distance [13], Jaro similarity [14], Dice [15], and Jaccard [16]. The corpus-based methods use information from the corpus to compute text similarity; this information can be either text features or co-occurrence probabilities. In recent studies, corpus-based approaches include three different measures: the bag-of-words model, distributed representation, and matrix decomposition method. The corpus-based methods mainly include bag-of-words [17], text frequency-inverse document frequency [18], Word2Vec [19], latent semantic analysis [20], and others. Semantic similarity determines the similarity between text and documents based on their meaning rather than character-by-character matching. Deep-structured semantic models [21] are typical models in this regard. Graph-based text similarities are mainly based on a knowledge-graph representation and a graph neural network representation. The graph structure better enables determining the similarity between nodes. Knowledge graphs [22] and graph neural networks [23] are the main methods for exploiting a graph structure.

Predicting the Future of AI in Health Care

Some previous works have also discussed the application of AI in medicine and possible future directions. One is integrative analysis [24], where data from different modalities can describe various aspects of a health problem. By mining these heterogeneous data in an integrated way, holistic and comprehensive insight into health can be obtained. In recent years, there has been a growing number of studies and initiatives related to AI in health, integrating different aspects of clinical data and linking drug development to clinical data. AI for precision medicine [25] represents another promising combination of AI and medicine, which assists in solving the most complex problems in personalized care. For example, AI in microscopic diagnostics [26] can improve the work of pathologists and may even gradually replace their work.

In this study, we used language models to measure the relationship between keywords, which can subsequently assist in building aggregation models and using adjacent keywords. Specifically, we propose a background-enhanced prediction method for constructing data for prediction using adjacent keywords, which refer to matrices adjacent to a 2D correlation matrix constructed using a clustering algorithm. This approach allows regression models to learn better and more accurately predict the relationships between keywords. We applied this approach to predict the future trend of AI technologies used in different biomedical domains based on past trends of related

technologies and biomedical domains. We further compared the prediction results to the patterns of current trends to evaluate the reliability of the prediction.

Methods

Data Sets

The data sets used in this study were obtained from the National Institutes of Health PubMed and PMC collections, with measures taken to avoid duplication by utilizing unique identifiers.

The corpus utilized in this study consists of three parts: (1) 114,266 abstracts and 49,126 full texts from PubMed and PMC obtained by searching keywords such as “machine learning,” “data mining,” “artificial intelligence,” “deep learning,” and “classifier” in the Title/Abstract field; (2) 61,382 full-text papers from PMC obtained by searching keywords such as “machine learning,” “data mining,” “artificial intelligence,” and “deep learning” in all fields, serving as a complement to the previous part; and (3) 2,507,391 full-text papers retrieved from the PubMed Central Open Access section with no keyword filtering to capture a comprehensive understanding of the biomedical field.

Due to permission restrictions, full-text access was limited for some papers. The full text of the papers primarily served for training, with the core of our experiments lying in the analysis and model prediction based on the abstracts.

Language Model

We utilized the word-embedding model Word2Vec as our language model owing to its advantages of efficiency and robustness among other available options [27].

Word embedding is a method of transforming a single word into a digital representation that captures various features of the word within a text, such as semantic relationships, definitions, and contexts. These digital representations can be used to identify similarities or dissimilarities between words.

To feed text data into a machine learning model, the text must be converted into an embedding. A simple method to achieve this involves “hot-coding” the text data, where each vector is mapped to a category. However, such simple embeddings have limitations as they do not capture the features of the words and can be large depending on the corpus size.

The effectiveness of Word2Vec is derived from its ability to combine vectors of similar words, leading to reliable estimates of word meaning based on their frequency in the corpus. This results in associations with other words, such as similar embedding vectors of “king” and “queen.” Algebraic operations on word embeddings can also provide approximations of word similarity, such as obtaining the vector for “queen” by subtracting the vector for “man” from the vector for “king” and adding the vector for “woman.” The cosine similarity measure is used to compare the similarity of two words, which is calculated according to the following formula:

$$\cos(x,y)=x \cdot y / |x| \times |y|$$

To improve the suitability of the original corpus for our language model, we performed extensive preprocessing to address any noise that may impact the model's effectiveness. This included removing all numeric and nonalphanumeric characters, except for the special character “-,” which is often used to link multiple words and create unique phrases. Additionally, to enhance the word vectors of biomedical- and AI-related keywords, we transformed multiword keywords in 114,266 abstracts into single tokens by merging them; for example, “machine learning” was merged into “machine+learning.”

The selection of hyperparameters was based on the available computational resources and the training corpus size. Our Word2Vec model had 300 dimensions and a window size of 5. Our computational device is a cluster with 384 GB of memory and 16 CPU cores. The Word2Vec model was trained sequentially on the three data sets, with the entire training process taking approximately 72 hours.

Background-Enhanced Prediction

Technology tends to be heavily studied in similar areas of research. Conversely, technology and its similar variants may be very popular in the same field. For example, techniques used for one type of cancer may also be relevant to other types of cancer, and various artificial neural models can all be applied in the field of medical image recognition. Our model was developed to predict future research trends based on direct relationships between technologies and fields and related technologies and fields. More specifically, we extracted the top 500 most frequent AI terms and the top 1000 most frequent biomedical fields from the 114,266 abstracts. To distinguish AI terms from biomedical terms, we adopted a simple classifier.

We obtained approximately 47,000 biomedical phrases from Medical Subject Headings and approximately 700 AI algorithms from Wikipedia. We used the average cosine similarity of each keyword and all terms in the two-word sets to predict whether the keyword should belong to the biomedical or AI domain. Next, Word2vec was used to obtain embeddings from each word. After converting all words into embeddings using Word2vec, we applied agglomerative clustering [28] to classify all the keywords according to their embeddings. Agglomerative clustering is a bottom-up clustering process. Initially, each input object forms its cluster. In each subsequent step, the two “closest” clusters are merged until only one remains. In our case, words with similar meanings will be grouped. Such a hierarchy is useful in many applications, and we provide the resulting tree diagram next to the corresponding heat map to best visualize the relationships between the surrounding categories.

Figure 1 depicts the co-occurrence frequency of biomedical and AI keywords. For regression prediction, we utilized not only the data from the orange part (information held by the keyword) but also from the green part (information held by the word neighboring the keyword). This inclusion provides a richer context, offering models that include more relevant information to learn from. The number 4 in the orange cell indicates the number of co-occurrences of “neural network” and “cancer,” which we not only used as input to predict the number of future co-occurrences of the terms “neural network” and “cancer” but also added the number of co-occurrences in the green section, $5+3+5+3+4+7+5+4$, to obtain a more comprehensive prediction using the neighboring information.

Figure 1. Co-occurrence frequency table of biomedical- and artificial intelligence-related keywords. Each number represents the number of co-occurrences of a given artificial intelligence model and biomedical term. The orange part represents the information held by the keyword and the green part represents the information held by the keyword's neighbors. CNN, convolutional neural network; LSTM: long short-term memory; MLP, multilayer perceptron; NN: neural network; RNN, recurrent neural network.

	Heart	Protein	Cancer	Tumor	Gene
LSTM	1	4	8	6	4
RNN	4	5	3	5	3
NN	5	3	4	4	7
CNN	3	7	5	4	6
MLP	3	9	6	5	6

Regression Model

The inputs and outputs of the regression model represent the co-occurrence frequency of biomedical and AI keywords in previous years and the co-occurrence frequency of future

biomedical and AI keywords obtained by prediction. Due to the limited number of AI-related papers from 1970 to 2000, we used semiannual statistics for January 2000 to December 2021 in our analysis. We incorporated each semiannual data set into a training and testing prediction model. Our model uses a small

After encoding words using Word2Vec, each word becomes a corresponding embedding. To evaluate the quality of the generated embeddings, we employed t-distributed stochastic neighbor embedding (t-SNE) [33], a technique for visualizing high-dimensional data by projecting it onto a 2D map. The t-SNE plots in Figures 3 and 4 reveal that the word embeddings

obtained by Word2Vec do allow words with similar meanings to be close together in the embedding space. Figure 3 highlights the vector positions of cancer-related keywords in 2D space, while Figure 4 shows the positions of classifier-related keywords.

Figure 3. Biomedical keywords in a t-distributed stochastic neighbor embedding plot.

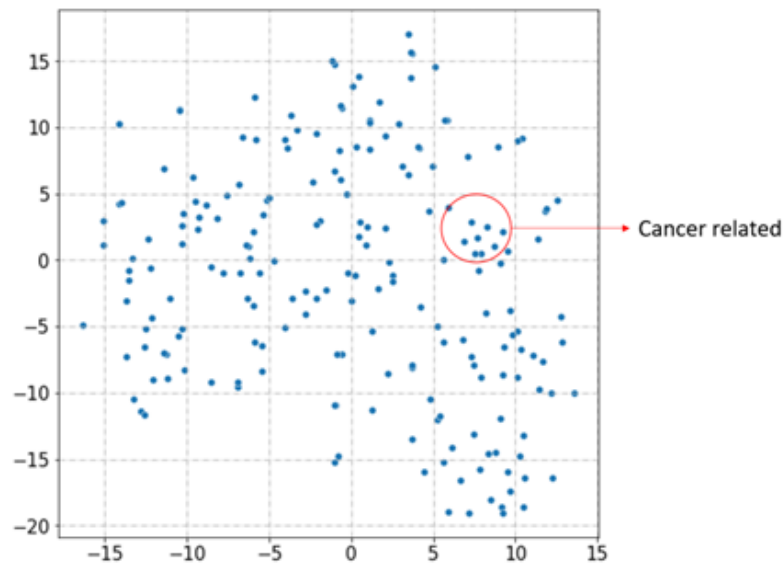
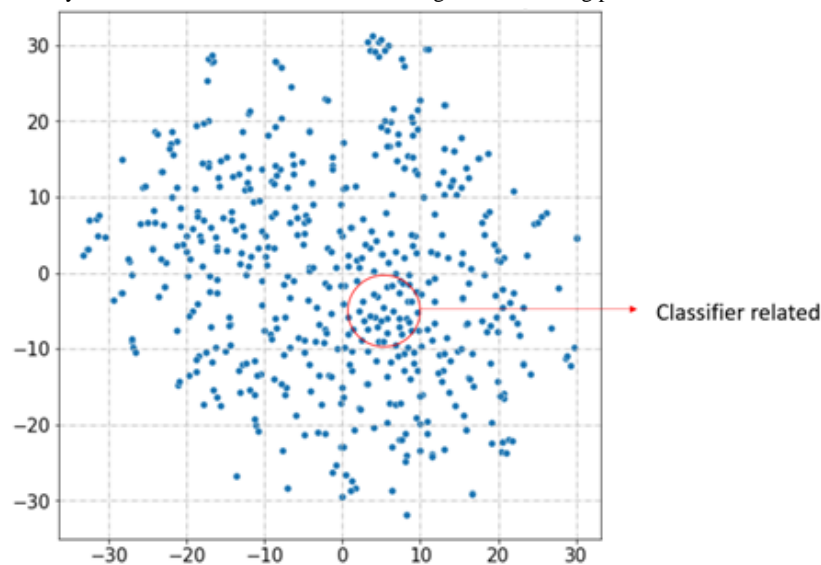


Figure 4. Artificial intelligence keywords in the t-distributed stochastic neighbor embedding plot.



Future Trend Prediction

Figure 5 illustrates the average R^2 values of all predicted and actual results from July 2002 to December 2021, with different window sizes of 1, 3, 5, 7, and 9. From Figure 5, we can also see that the elastic net model provided the best results when the window size was equal to 9, whereas some other models worked best when the window size was equal to 3.

Since our model relies on the previous year's heatmap as a feature, to predict a longer time horizon, we iteratively ran our model using the predicted heatmap of cycle x to predict the heatmap of cycle $x+1$. As shown in Figure 6, although the R^2

value decreased during the 5-year prediction, it was still relatively high. We also provide a 100×200 demonstration to visualize the prediction results in Figures 7-10. These heatmaps, like those in Figure 2, are also used to show the frequency of co-occurrence between the keywords of AI technology and biomedicine. Figure 8 depicts the original publications that were recorded between July and December 2021, while Figure 9 represents the predicted publications for the same time period. To effectively showcase the disparity between the actual and projected outcomes, a heatmap was generated using both the original and predicted heatmaps. This comparison is visually presented in Figure 10, allowing for a clear and easily understandable differentiation between the two sets of data.

Figure 5. Mean R-square values obtained by forecasting in half-yearly intervals from July 2002 to December 2021 under different window sizes for different methods. SVR: support vector regression.

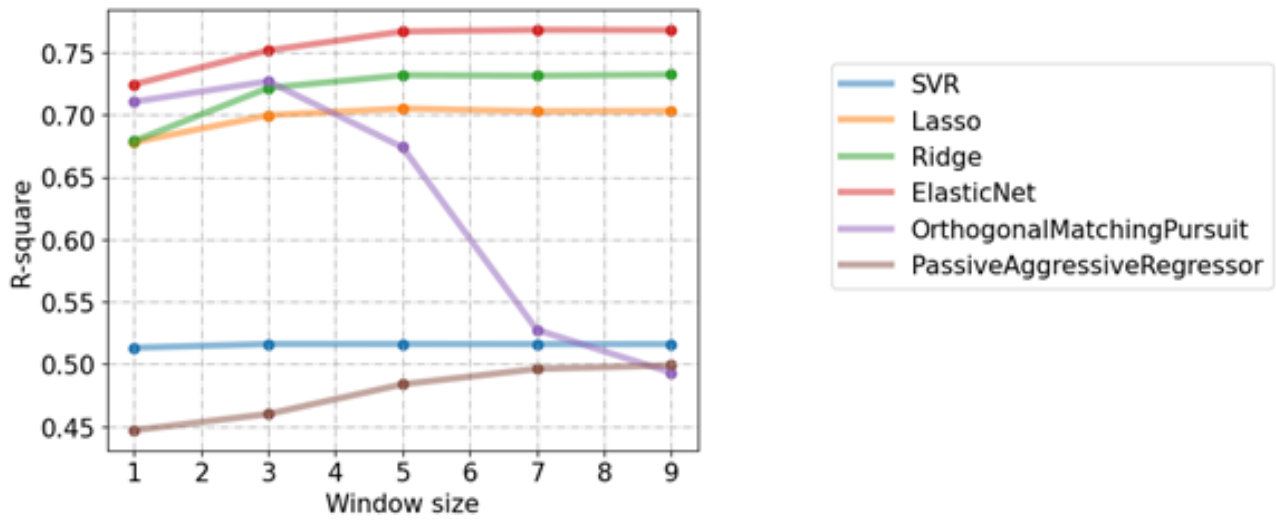


Figure 6. Line graph of the forecast results for each half year from 2002 to 2021. The model used was elastic net with a 9x9 window size, as this resulted in the best prediction (R-square value).

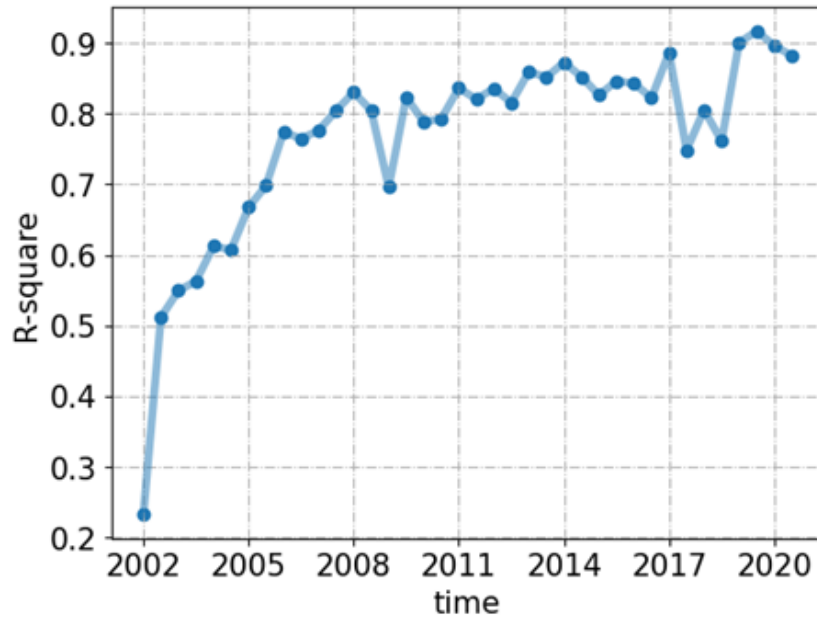


Figure 7. The predictions are iterated in half-year increments from July 2014 to December 2021, and the data obtained from the predictions are used as the data set for the subsequent prediction models for training. The horizontal axis is time and the vertical axis is the R-square value. A higher resolution version of this figure is available in [Multimedia Appendix 2](#).

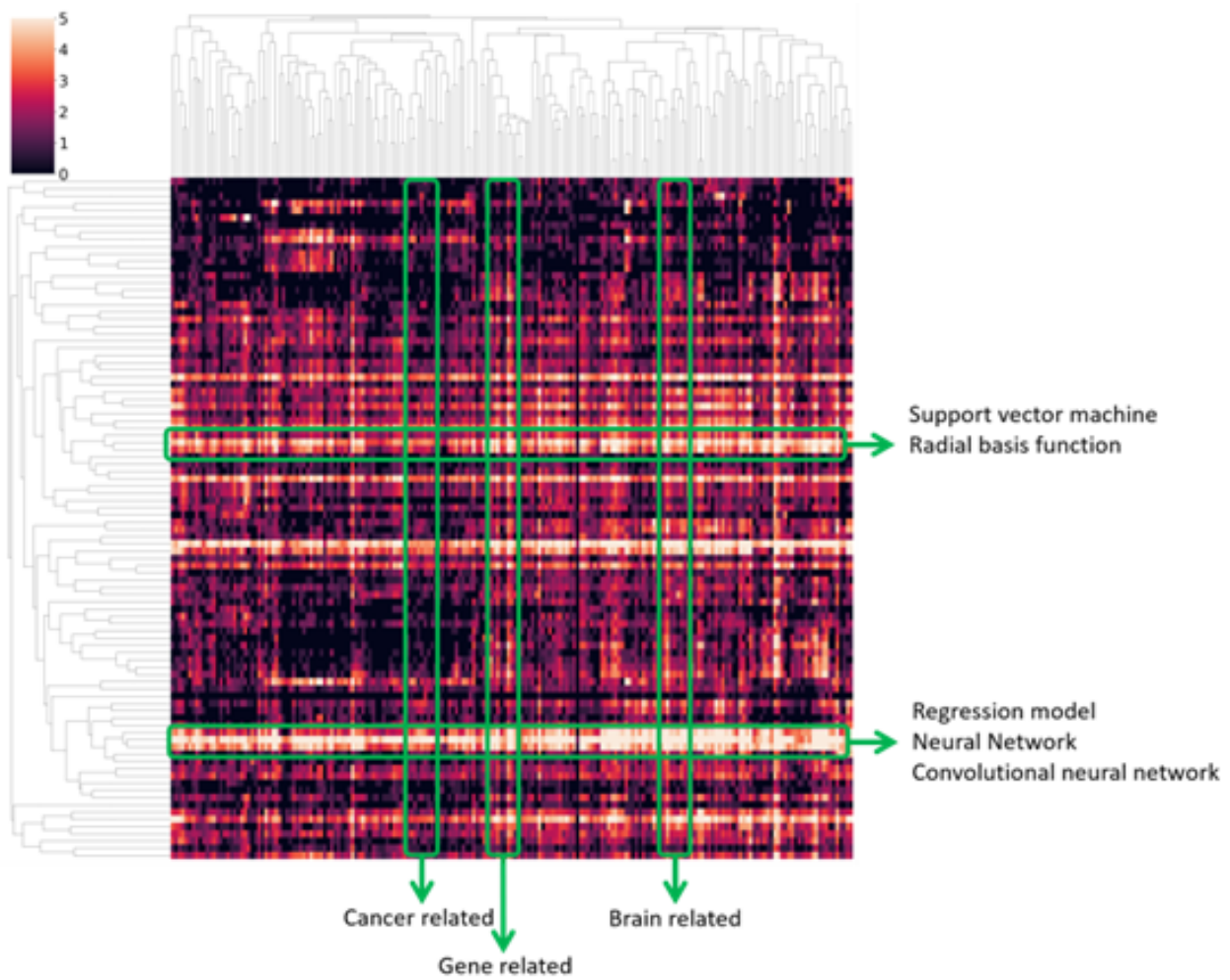


Figure 8. Heatmap from July to December 2021 for the actual intersection of artificial intelligence (AI) technology and biomedical field applications. The horizontal axis is the keywords in the medical field and the vertical axis is the keywords in AI technology. A higher resolution version of this figure is available in [Multimedia Appendix 3](#).

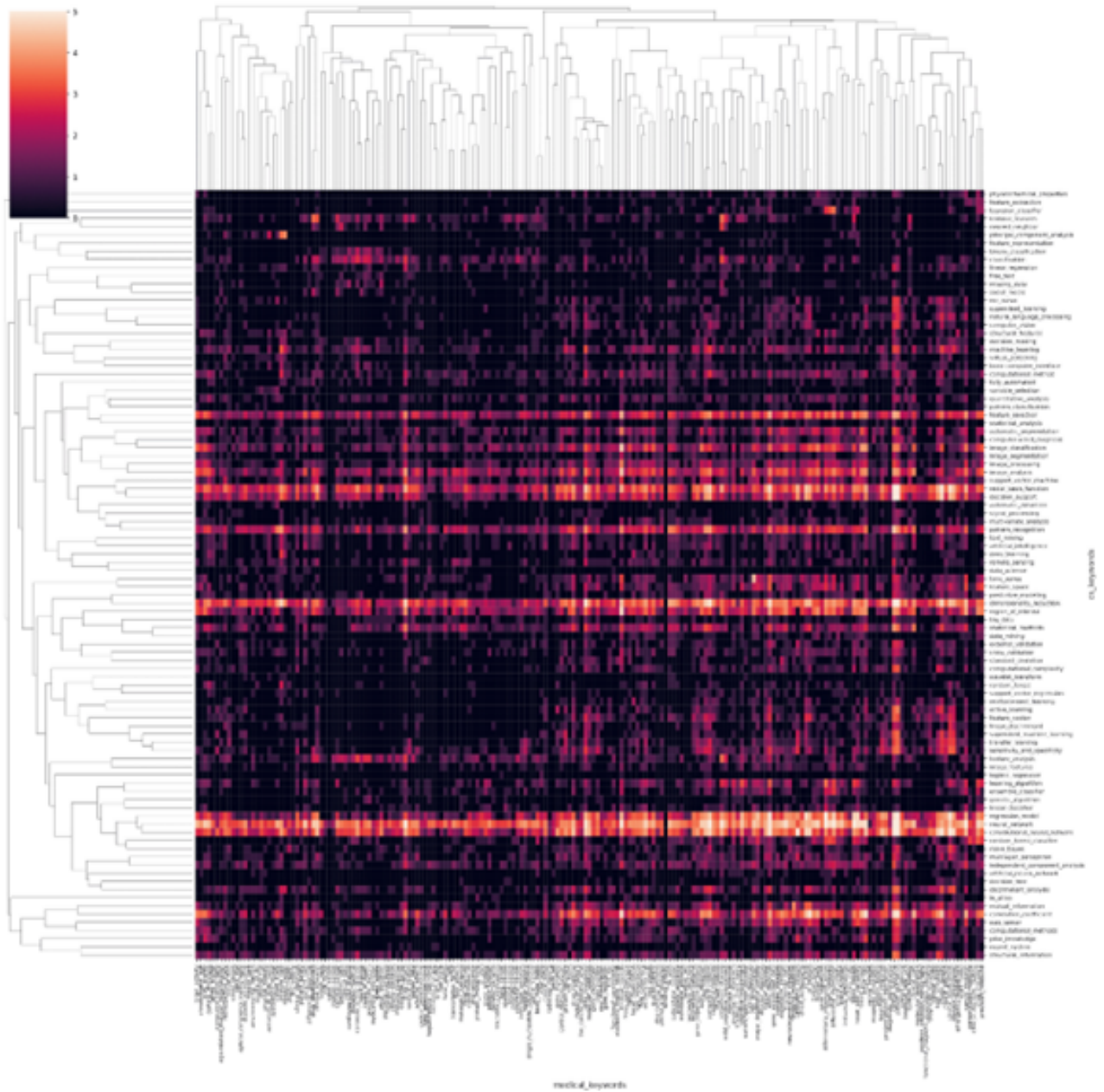


Figure 9. Predicted heat map of the intersection of artificial intelligence (AI) technology and biomedical field applications from July to December 2021. The horizontal axis is the keywords in the medical field and the vertical axis is the keywords in AI technology. A higher resolution version of this figure is available in [Multimedia Appendix 4](#).

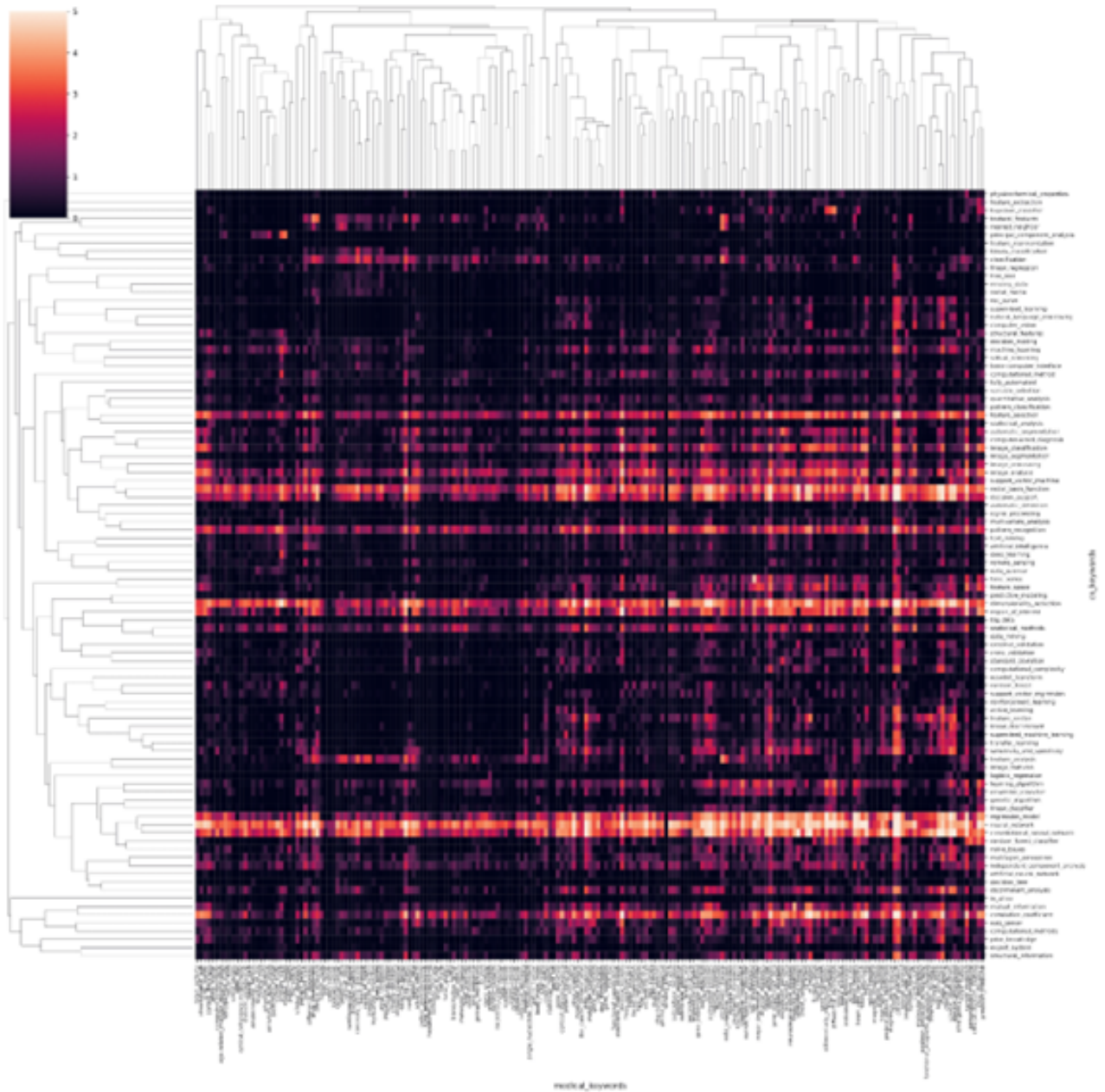
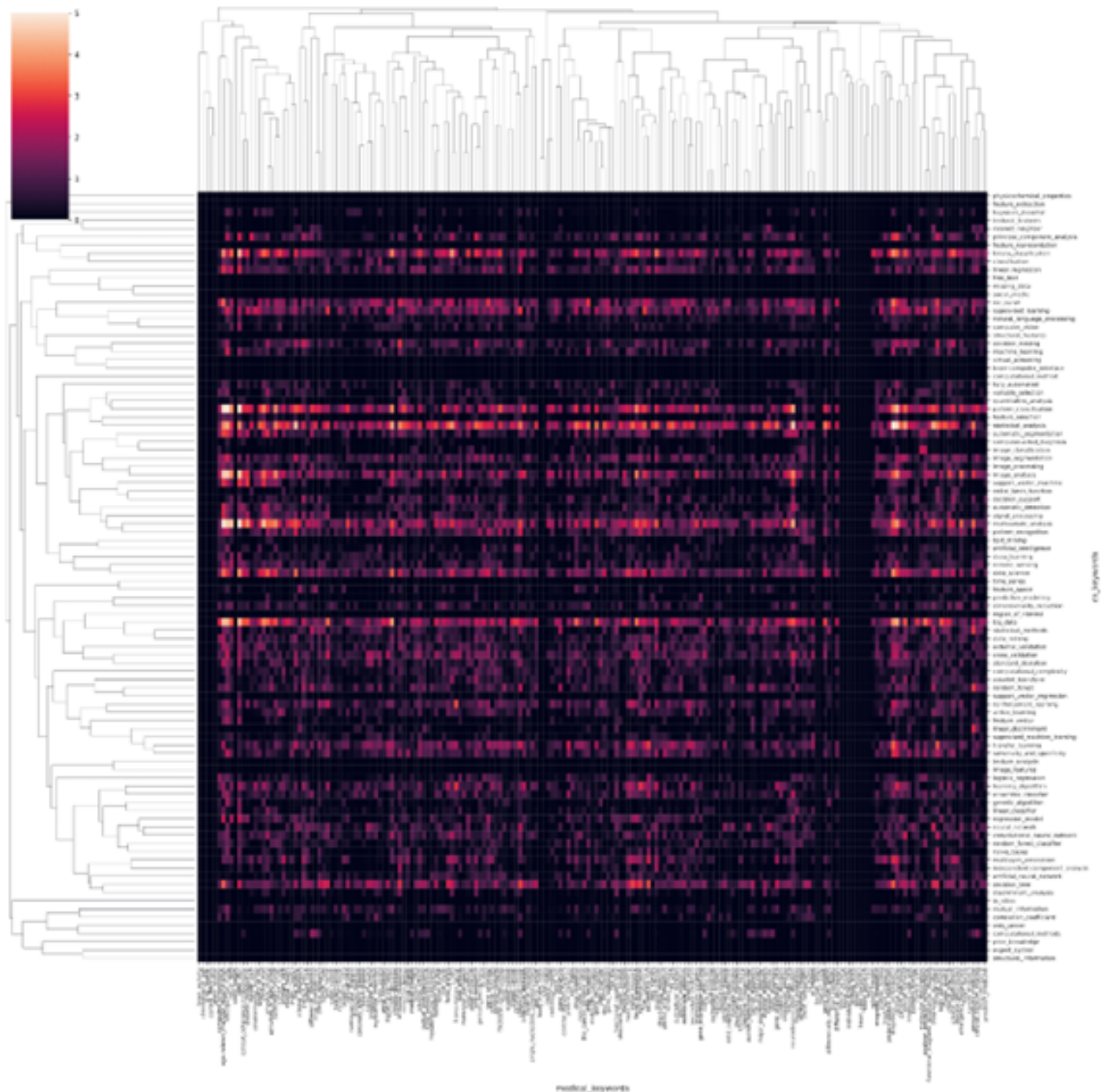


Figure 10. Heatmap drawn from the difference between the predicted and actual heatmaps for July to December 2021 (Figures 9 and 8, respectively) representing the intersection of artificial intelligence (AI) technology and biomedical field applications. The horizontal axis is the keywords in the medical field and the vertical axis is the keywords in AI technology. A higher resolution version of this figure is available in [Multimedia Appendix 5](#).



Co-occurrence Trend Analysis

The data obtained through statistical analysis indicated that the number of papers combining AI with biomedicine is increasing in spurts. From [Table 1](#), we can see which combinations between AI and biomedicine are the most popular. The field of genetics shows many combinations with various AI technologies, occupying 13 of the top 20 positions. Numerous papers on this topic highlight its popularity [34-36]. The combination of AI and protein ranked fourth, demonstrating that protein analysis is a very suitable field for the use of machines. Cancer and tumors are currently the main challenges in biomedicine, and

their combination with AI is also a popular topic at present. In these biomedical fields, machine learning is the AI technology with the highest number of applications. Although deep learning and neural networks are trendy, traditional methods such as vector automata and random forests are still the main choices in biomedical fields. Many fundamental concepts of AI are also included in this ranking, such as classification, regression, cross-validation, feature extraction, receiver operating characteristic, and others. Overall, this analysis shows that AI has become a key technology in the biomedical field and requires the proficiency of biomedical scientists.

Table 1. The top 20 combinations of artificial intelligence (AI) technologies and biomedical fields that have appeared in the literature in the last 5 years.

Rank	AI technology, biomedical field	Proportion of publications, %
1	machine learning, gene	1.650
2	classification, gene	1.038
3	neural network, gene	0.634
4	deep learning, gene	0.453
5	support vector machine, gene	0.447
6	machine learning, protein	0.404
7	regression, gene	0.402
8	learning algorithm, gene	0.385
9	machine learning, cancer	0.380
10	classification, cancer	0.351
11	random forest, gene	0.331
12	artificial intelligence, gene	0.249
13	convolution neural network, gene	0.241
14	cross-validation, gene	0.219
15	feature selection, gene	0.191
16	neural network, cancer	0.176
17	classification, tumor	0.172
18	supervised learning, gene	0.171
19	machine learning, tumor	0.171
20	receiver operating characteristic, gene	0.169

Since many combinations between AI and biomedicine have a very small contribution or are nonexistent, some are not meaningful; therefore, we set a reasonable threshold to filter such combinations, avoiding the situation where the original minimal combination grows by a considerable percentage with little growth so that the table showing the trend of changes is more meaningful. From [Table 2](#), we can see the very rapid growth of cases combining AI and biomedicine in the last 5 years. This is because genes, proteins, oncology, and many other fields are growing rapidly, and core medical testing technology such as magnetic resonance imaging is compatible with AI.

We used the best model from our proposed methodology to forecast the trends in AI technology and biomedicine over the next 5 years. The prediction results for the contributions of each combination and their growth are shown in [Table 3](#) and [Table](#)

[4](#), respectively. The regression results were rounded for brevity of presentation in the tables. We can use these predicted results to provide an outlook on the future development of AI in biomedicine. From the point of view of AI technologies, standard techniques such as deep learning, machine learning, and neural networks still dominate. Traditional machine learning methods such as random forest and support vector machine are outside the top 20 prediction results. Deep learning will gradually become the mainstream AI technology combined with biomedicine [37]. From a biomedical perspective, genetics will continue to dominate. At the same time, studies focusing on proteins and tumors will leave the top 20 and be replaced by early diagnostics, tomography, and other detection technologies. These are certain areas that are well suited to incorporate AI technology.

Table 2. The 20 most rapidly growing combinations of artificial intelligence (AI) technologies and biomedical fields in the last 5 years.

Rank	AI technology, biomedical field	Growth, %
1	electronic, health records	1054.545
2	electronic health records, electronic health	1054.545
3	machine learning, electronic health record	1033.333
4	machine learning, health care	820.000
5	machine learning, risk factor	816.667
6	machine learning, public health	735.000
7	neural network, gene	700.483
8	neural network, cancer	647.059
9	machine learning, tumor	619.697
10	image analysis, gene	613.333
11	machine learning, clinical trial	572.414
12	machine learning, clinical practice	566.667
13	decision making, gene	547.619
14	artificial intelligence, gene	511.111
15	random forest, cancer	493.617
16	machine learning, clinical data	487.179
17	electronic medical record, medical records	480.000
18	next generation sequencing, gene	467.647
19	random forest, tumor	466.667
20	machine learning, magnetic resonance	456.579

Table 3. The top 20 combinations of artificial intelligence (AI) technologies and biomedical fields that will emerge in the next 5 years.

Rank	AI technology, biomedical field	Predicted proportion of publications, %
1	machine learning, gene	2.331
2	artificial intelligence, early diagnosis	2.289
3	artificial intelligence, early detection	1.901
4	artificial intelligence, gene	1.487
5	neural network, gene	1.392
6	deep learning, computed tomography	1.288
7	artificial intelligence, systematic reviews	1.239
8	classification, gene	1.197
9	supervised learning, gene	1.188
10	generative adversarial network, gene	1.040
11	artificial intelligence, personalized treatment	0.881
12	machine learning, risk factors	0.659
13	deep learning, gene	0.633
14	artificial intelligence, systematic review	0.617
15	convolution neural network, gene	0.604
16	learning algorithm, gene	0.593
17	receiver operating characteristic, computed tomography scans	0.581
18	machine learning, medical records	0.578
19	machine learning, blood pressure	0.569
20	artificial intelligence, imaging modalities	0.554

Table 4. The top 20 rapidly growing combinations of artificial intelligence (AI) technology and biomedical fields in the next 5 years.

Rank	AI technology, biomedical field	Predicted growth, %
1	artificial intelligence, gene	2253.521
2	machine learning, risk factor	2184.491
3	cross-validation, gene	2164.150
4	receiver operating characteristic, gene	1504.581
5	learning algorithm, gene	1421.751
6	neural network, gene	1340.880
7	convolution neural network, gene	1296.067
8	classification, gene	1280.985
9	machine learning, gene	1261.342
10	classification, cancer	888.106
11	support vector machine, gene	791.807
12	neural network, cancer	665.430
13	artificial intelligence, cancer	621.627
14	deep learning, gene	502.318
15	classification, tumor	415.298
16	regression, gene	377.864
17	machine learning, protein	333.778
18	random forest, gene	322.787
19	deep learning, cancer	200.080
20	natural language processing, natural language	192.518

Discussion

Principal Findings

AI Technology Trends in Biomedicine

Our findings confirm that standard AI techniques, including deep learning, machine learning, and neural networks, continue to be the primary driving forces behind the integration of AI into biomedicine. However, it is noteworthy that generative adversarial networks (GANs) [38] are gaining prominence, particularly in the genetics field. GANs hold immense potential for applications in medical imaging and drug discovery owing to their ability to generate synthetic images across various modalities.

Evolution of Biomedical Research

The data also highlight the shifting landscape of biomedical research. While genetics remains dominant, areas such as proteins and tumors are gradually giving way to early diagnostics, tomography, and other detection technologies. These developments align with the suitability of these fields for AI integration, resulting in promising advancements in health care analysis and diagnostics.

Impact of AI on Health Care

As suggested by previous research [24], the future of AI in health care is promising. AI has the potential to enhance the accuracy of cancer diagnosis and prognosis beyond that of average statistical experts [39,40]. Furthermore, as AI

technology continues to advance, it will enable the resolution of more complex and specialized health care problems, further transforming the biomedical landscape.

Future Work

By utilizing keywords to filter medical papers that have applied AI techniques, we identified key connections and trends among them. The approach of using keywords aggregated based on text similarity performed well in the regression model. This approach is intuitive and leads to improved co-word analysis for trend prediction.

Fundamentally, incorporating peripheral information led to higher regression accuracy and more accurate predictions of future trends. Additionally, this approach also takes into account internal relationships within a class compared to previous methods. However, this also raises the question of how to best measure the degree of keyword association.

We made some simple assumptions that words with similar meanings would complement the information of the others. Specifically, considering only their own meanings tends to make the predictions one-sided, while having more reference information naturally makes the predictions more robust. This can be seen as a type of data augmentation. There are still many directions to explore regarding this approach. In future research, it may be possible to use different text similarity methods such as convolutional neural network, bidirectional encoder representations from transformers, and various regression models, where the reliability of text similarity determines

whether the information obtained from the surrounding context is valid. Additionally, different time spans for the prediction can be studied. Although this study focused on AI techniques in the biomedical field, the applicability of the proposed approach extends to any study involving co-word analysis.

Limitations

While our study provides valuable insights into the trends of AI technologies in the biomedical domain based on a comprehensive data set from PubMed, there are several limitations to consider. First, there is a limitation of the data source, since our study solely relies on PubMed as the primary source of articles, which might introduce a selection bias. There are numerous other databases and grey literature sources that were not considered, and their inclusion might have offered a more comprehensive view. Second, our study lacks external validity. Our findings, although significant in the context of our data set, require validation with real-world applications and events to check their external validity.

Acknowledgments

We are sincerely grateful to our advisor, Professor Soroush Vosoughi, for the invaluable guidance and support provided during the development of this paper. We would also like to thank our colleagues and the staff at Minds, Machines, and Society Lab for their helpful insights and assistance.

Data Availability

The code is available at GitHub [41].

Conflicts of Interest

None declared.

Multimedia Appendix 1

Higher resolution version of Figure 2.

[PDF File (Adobe PDF File), 427 KB - [ai_v2i1e45770_app1.pdf](#)]

Multimedia Appendix 2

Higher resolution version of Figure 7.

[PDF File (Adobe PDF File), 209 KB - [ai_v2i1e45770_app2.pdf](#)]

Multimedia Appendix 3

Higher resolution version of Figure 8.

[PDF File (Adobe PDF File), 346 KB - [ai_v2i1e45770_app3.pdf](#)]

Multimedia Appendix 4

Higher resolution version of Figure 9.

[PDF File (Adobe PDF File), 477 KB - [ai_v2i1e45770_app4.pdf](#)]

Multimedia Appendix 5

Higher resolution version of Figure 10.

[PDF File (Adobe PDF File), 421 KB - [ai_v2i1e45770_app5.pdf](#)]

References

1. Yu K, Beam AL, Kohane IS. Artificial intelligence in healthcare. *Nat Biomed Eng* 2018 Oct;2(10):719-731. [doi: [10.1038/s41551-018-0305-z](https://doi.org/10.1038/s41551-018-0305-z)] [Medline: [31015651](https://pubmed.ncbi.nlm.nih.gov/31015651/)]

2. Zemouri R, Zerhouni N, Racoceanu D. Deep learning in the biomedical applications: recent and future status. *Appl Sci* 2019 Apr 12;9(8):1526. [doi: [10.3390/app9081526](https://doi.org/10.3390/app9081526)]
3. Wei Y, Zhou J, Wang Y, Liu Y, Liu Q, Luo J, et al. *IEEE Trans Biomed Circuits Syst* 2020 Apr;14(2):145-163. [doi: [10.1109/TBCAS.2020.2974154](https://doi.org/10.1109/TBCAS.2020.2974154)] [Medline: [32078560](https://pubmed.ncbi.nlm.nih.gov/32078560/)]
4. Bali J, Garg R, Bali RT. Artificial intelligence (AI) in healthcare and biomedical research: Why a strong computational/AI bioethics framework is required? *Indian J Ophthalmol* 2019 Jan;67(1):3-6 [FREE Full text] [doi: [10.4103/ijo.IJO_1292_18](https://doi.org/10.4103/ijo.IJO_1292_18)] [Medline: [30574881](https://pubmed.ncbi.nlm.nih.gov/30574881/)]
5. Donthu N, Kumar S, Mukherjee D, Pandey N, Lim WM. How to conduct a bibliometric analysis: an overview and guidelines. *J Bus Res* 2021 Sep;133:285-296. [doi: [10.1016/j.jbusres.2021.04.070](https://doi.org/10.1016/j.jbusres.2021.04.070)]
6. He Q. Knowledge discovery through co-word analysis. *Library Trends* 1999;48(1):133-159 [FREE Full text]
7. Wang J, Dong Y. Measurement of text similarity: a survey. *Information* 2020 Aug 31;11(9):421. [doi: [10.3390/info11090421](https://doi.org/10.3390/info11090421)]
8. Deza M, Deza E. *Encyclopedia of distances*. Berlin, Heidelberg: Springer; 2009.
9. Manning C, Schütze H. *Foundations of statistical natural language processing*. Cambridge, MA: MIT Press; 1999.
10. Kullback S, Leibler RA. On information and sufficiency. *Ann Math Statist* 1951 Mar;22(1):79-86. [doi: [10.1214/aoms/1177729694](https://doi.org/10.1214/aoms/1177729694)]
11. Weng L. From GAN to WGAN. arXiv. 2019. URL: <http://arxiv.org/abs/1904.08994> [accessed 2023-11-26]
12. Irving R, Fraser C. Two algorithms for the longest common subsequence of three (or more) strings. In: Apostolico A, Crochemore M, Galil Z, Manber U, editors. *Combinatorial Pattern Matching Combinatorial Pattern Matching*. CPM 1992. Lecture Notes in Computer Science, vol 644. Berlin, Heidelberg: Springer; 1992:214-229.
13. Damerau FJ. A technique for computer detection and correction of spelling errors. *Commun ACM* 1964 Mar;7(3):171-176. [doi: [10.1145/363958.363994](https://doi.org/10.1145/363958.363994)]
14. ERIC: String Comparator Metrics and Enhanced Decision Rules in the Fellegi-Sunter Model of Record Linkage. URL: <https://eric.ed.gov/?id=ED325505> [accessed 2023-11-26]
15. Dice LR. Measures of the amount of ecologic association between species. *Ecology* 1945 Jul;26(3):297-302. [doi: [10.2307/1932409](https://doi.org/10.2307/1932409)]
16. Jaccard P. The distribution of the flora in the alpine zone. *New Phytologist* 1912;11(2):37-50 [FREE Full text]
17. Salton G, Buckley C. Term weighting approaches in automatic text retrieval. *Inf Process Manag* 1988;24(5):513-523. [doi: [10.1016/0306-4573\(88\)90021-0](https://doi.org/10.1016/0306-4573(88)90021-0)]
18. Robertson S, Walker S. Some simple effective approximations to the 2-Poisson model for probabilistic weighted retrieval. 1994 Presented at: Seventeenth Annual International ACM-SIGIR Conference on Research and Development in Information Retrieval (SIGIR '94); July 3-6, 1994; Dublin, Ireland. [doi: [10.1007/978-1-4471-2099-5_24](https://doi.org/10.1007/978-1-4471-2099-5_24)]
19. Rong X. word2vec parameter learning explained. arXiv. 2016. URL: <http://arxiv.org/abs/1411.2738> [accessed 2023-11-26]
20. Deerwester S, Dumais ST, Furnas GW, Landauer TK, Harshman R. Indexing by latent semantic analysis. *J Am Soc Inf Sci* 1990 Sep;41(6):391-407. [doi: [10.1002/\(sici\)1097-4571\(199009\)41:6<391::aid-asi1>3.0.co;2-9](https://doi.org/10.1002/(sici)1097-4571(199009)41:6<391::aid-asi1>3.0.co;2-9)]
21. Shen Y, He X, Gao J, Deng L, Mesnil G. A latent semantic model with convolutional-pooling structure for information retrieval. 2014 Presented at: 23rd ACM International Conference on Conference on Information Knowledge Management; November 3-7, 2014; Shanghai, China. [doi: [10.1145/2661829.2661935](https://doi.org/10.1145/2661829.2661935)]
22. Chen X, Jia S, Xiang Y. A review: knowledge reasoning over knowledge graph. *Expert Syst Appl* 2020 Mar;141:112948. [doi: [10.1016/j.eswa.2019.112948](https://doi.org/10.1016/j.eswa.2019.112948)]
23. Zhou J, Cui G, Hu S, Zhang Z, Yang C, Liu Z, et al. Graph neural networks: a review of methods and applications. *AI Open* 2020;1:57-81. [doi: [10.1016/j.aiopen.2021.01.001](https://doi.org/10.1016/j.aiopen.2021.01.001)]
24. Wang F, Preininger A. AI in health: state of the art, challenges, and future directions. *Yearb Med Inform* 2019 Aug;28(1):16-26 [FREE Full text] [doi: [10.1055/s-0039-1677908](https://doi.org/10.1055/s-0039-1677908)] [Medline: [31419814](https://pubmed.ncbi.nlm.nih.gov/31419814/)]
25. Johnson KB, Wei W, Weeraratne D, Frisse ME, Misulis K, Rhee K, et al. Precision medicine, AI, and the future of personalized health care. *Clin Transl Sci* 2021 Jan;14(1):86-93 [FREE Full text] [doi: [10.1111/cts.12884](https://doi.org/10.1111/cts.12884)] [Medline: [32961010](https://pubmed.ncbi.nlm.nih.gov/32961010/)]
26. Ahmad Z, Rahim S, Zubair M, Abdul-Ghafar J. Artificial intelligence (AI) in medicine, current applications and future role with special emphasis on its potential and promise in pathology: present and future impact, obstacles including costs and acceptance among pathologists, practical and philosophical considerations. A comprehensive review. *Diagn Pathol* 2021 Mar 17;16(1):24 [FREE Full text] [doi: [10.1186/s13000-021-01085-4](https://doi.org/10.1186/s13000-021-01085-4)] [Medline: [33731170](https://pubmed.ncbi.nlm.nih.gov/33731170/)]
27. Hu K, Luo Q, Qi K, Yang S, Mao J, Fu X, et al. Understanding the topic evolution of scientific literatures like an evolving city: using Google Word2Vec model and spatial autocorrelation analysis. *Inf Processing Manag* 2019 Jul;56(4):1185-1203. [doi: [10.1016/j.ipm.2019.02.014](https://doi.org/10.1016/j.ipm.2019.02.014)]
28. Ackermann MR, Blömer J, Kuntze D, Sohler C. Analysis of agglomerative clustering. *Algorithmica* 2012 Dec 12;69(1):184-215. [doi: [10.1007/s00453-012-9717-4](https://doi.org/10.1007/s00453-012-9717-4)]
29. Xiao R, Cui X, Qiao H, Zheng X, Zhang Y. Early diagnosis model of Alzheimer's Disease based on sparse logistic regression. *Multimed Tools Appl* 2020 Sep 25;80(3):3969-3980. [doi: [10.1007/s11042-020-09738-0](https://doi.org/10.1007/s11042-020-09738-0)]

30. Zarei A, Asl BM. Automatic seizure detection using orthogonal matching pursuit, discrete wavelet transform, and entropy based features of EEG signals. *Comput Biol Med* 2021 Apr;131:104250. [doi: [10.1016/j.combiomed.2021.104250](https://doi.org/10.1016/j.combiomed.2021.104250)] [Medline: [33578071](https://pubmed.ncbi.nlm.nih.gov/33578071/)]
31. Malki Z, Atlam E, Hassanien AE, Dagnew G, Elhosseini MA, Gad I. Association between weather data and COVID-19 pandemic predicting mortality rate: machine learning approaches. *Chaos Solitons Fractals* 2020 Sep;138:110137 [FREE Full text] [doi: [10.1016/j.chaos.2020.110137](https://doi.org/10.1016/j.chaos.2020.110137)] [Medline: [32834583](https://pubmed.ncbi.nlm.nih.gov/32834583/)]
32. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B. Scikit-learn: machine learning in Python. *J Machine Learn Res* 2011;12:2825-2830 [FREE Full text]
33. van der Maaten L, Hinton G. Visualizing data using t-SNE. *J Mach Learn Res* 2008;9:2579-2605 [FREE Full text]
34. Abdulkader DM, Abdulazeez AM, Zeebaree D. Machine learning supervised algorithms of gene selection: a review. *Technology Reports of Kansai University* 2020;62(3):233-244 [FREE Full text] [doi: [10.1109/icset53708.2021.9612526](https://doi.org/10.1109/icset53708.2021.9612526)]
35. Mahood EH, Kruse LH, Moghe GD. Machine learning: a powerful tool for gene function prediction in plants. *Appl Plant Sci* 2020 Jul;8(7):e11376 [FREE Full text] [doi: [10.1002/aps3.11376](https://doi.org/10.1002/aps3.11376)] [Medline: [32765975](https://pubmed.ncbi.nlm.nih.gov/32765975/)]
36. Mochida K, Koda S, Inoue K, Nishii R. Statistical and machine learning approaches to predict gene regulatory networks from transcriptome datasets. *Front Plant Sci* 2018;9:1770 [FREE Full text] [doi: [10.3389/fpls.2018.01770](https://doi.org/10.3389/fpls.2018.01770)] [Medline: [30555503](https://pubmed.ncbi.nlm.nih.gov/30555503/)]
37. Zhang L, Tan J, Han D, Zhu H. From machine learning to deep learning: progress in machine intelligence for rational drug discovery. *Drug Discov Today* 2017 Nov;22(11):1680-1685. [doi: [10.1016/j.drudis.2017.08.010](https://doi.org/10.1016/j.drudis.2017.08.010)] [Medline: [28881183](https://pubmed.ncbi.nlm.nih.gov/28881183/)]
38. Goodfellow I, Pouget-Abadie J, Mirza M, Xu B, Warde-Farley D, Ozair S, et al. Generative adversarial networks. *Commun ACM* 2020 Oct 22;63(11):139-144. [doi: [10.1145/3422622](https://doi.org/10.1145/3422622)]
39. Huang S, Yang J, Fong S, Zhao Q. Artificial intelligence in cancer diagnosis and prognosis: opportunities and challenges. *Cancer Lett* 2020 Feb 28;471:61-71. [doi: [10.1016/j.canlet.2019.12.007](https://doi.org/10.1016/j.canlet.2019.12.007)] [Medline: [31830558](https://pubmed.ncbi.nlm.nih.gov/31830558/)]
40. Belić M, Bobić V, Badža M, Šolaja N, Đurić-Jovičić M, Kostić VS. Artificial intelligence for assisting diagnostics and assessment of Parkinson's disease-A review. *Clin Neurol Neurosurg* 2019 Sep;184:105442. [doi: [10.1016/j.clineuro.2019.105442](https://doi.org/10.1016/j.clineuro.2019.105442)] [Medline: [31351213](https://pubmed.ncbi.nlm.nih.gov/31351213/)]
41. Code used in this study. GitHub. URL: https://github.com/jiashenggu/ai_in_bio [accessed 2023-11-26]

Abbreviations

AI: artificial intelligence

GAN: generative adversarial network

t-SNE: t-distributed stochastic neighbor embedding

Edited by K El Emam, B Malin; submitted 16.01.23; peer-reviewed by L Huang, JA Benítez-Andrades, D Kohen; comments to author 19.05.23; revised version received 11.06.23; accepted 29.10.23; published 19.12.23.

Please cite as:

Gu J, Gao C, Wang L

The Evolution of Artificial Intelligence in Biomedicine: Bibliometric Analysis

JMIR AI 2023;2:e45770

URL: <https://ai.jmir.org/2023/1/e45770>

doi: [10.2196/45770](https://doi.org/10.2196/45770)

PMID: [38875563](https://pubmed.ncbi.nlm.nih.gov/38875563/)

©Jiasheng Gu, Chongyang Gao, Lili Wang. Originally published in JMIR AI (<https://ai.jmir.org>), 19.12.2023. This is an open-access article distributed under the terms of the Creative Commons Attribution License (<https://creativecommons.org/licenses/by/4.0/>), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR AI, is properly cited. The complete bibliographic information, a link to the original publication on <https://www.ai.jmir.org/>, as well as this copyright and license information must be included.

Original Paper

Association Between Online Reviews of Substance Use Disorder Treatment Facilities and Drug-Induced Mortality Rates: Cross-Sectional Analysis

Matthew P Abrams^{1,2,3}, MD; Raina M Merchant^{1,2,4}, MD, MSc; Zachary F Meisel^{2,4,5}, MD, MPH, MSc; Arthur P Pelullo¹, MS, MA; Sharath Chandra Guntuku^{1,4,6}, PhD; Anish K Agarwal^{1,2,4}, MD, MPH, MS

¹Center for Digital Health, University of Pennsylvania Perelman School of Medicine, Philadelphia, PA, United States

²Center for Emergency Care Policy and Research, University of Pennsylvania Perelman School of Medicine, Philadelphia, PA, United States

³Department of Psychiatry, University of California San Diego, San Diego, CA, United States

⁴Leonard Davis Institute of Health Economics, University of Pennsylvania, Philadelphia, PA, United States

⁵Penn Injury Science Center, University of Pennsylvania, Philadelphia, PA, United States

⁶Department of Computer and Information Science, University of Pennsylvania, Philadelphia, PA, United States

Corresponding Author:

Matthew P Abrams, MD

Center for Emergency Care Policy and Research
University of Pennsylvania Perelman School of Medicine
4th Floor, Blockley Hall, 423 Guardian Drive
Philadelphia, PA, 19104-6021

United States

Phone: 1 000 000 0000

Email: matthew.abrams@pennmedicine.upenn.edu

Abstract

Background: Drug-induced mortality across the United States has continued to rise. To date, there are limited measures to evaluate patient preferences and priorities regarding substance use disorder (SUD) treatment, and many patients do not have access to evidence-based treatment options. Patients and their families seeking SUD treatment may begin their search for an SUD treatment facility online, where they can find information about individual facilities, as well as a summary of patient-generated web-based reviews via popular platforms such as Google or Yelp. Web-based reviews of health care facilities may reflect information about factors associated with positive or negative patient satisfaction. The association between patient satisfaction with SUD treatment and drug-induced mortality is not well understood.

Objective: The objective of this study was to examine the association between online review content of SUD treatment facilities and drug-induced state mortality.

Methods: A cross-sectional analysis of online reviews and ratings of Substance Abuse and Mental Health Services Administration (SAMHSA)-designated SUD treatment facilities listed between September 2005 and October 2021 was conducted. The primary outcomes were (1) mean online rating of SUD treatment facilities from 1 star (worst) to 5 stars (best) and (2) average drug-induced mortality rates from the Centers for Disease Control and Prevention (CDC) WONDER Database (2006-2019). Clusters of words with differential frequencies within reviews were identified. A 3-level linear model was used to estimate the association between online review ratings and drug-induced mortality.

Results: A total of 589 SAMHSA-designated facilities (n=9597 reviews) were included in this study. Drug-induced mortality was compared with the average. Approximately half (24/47, 51%) of states had below average (“low”) mortality rates (mean 13.40, SD 2.45 deaths per 100,000 people), and half (23/47, 49%) had above average (“high”) drug-induced mortality rates (mean 21.92, SD 3.69 deaths per 100,000 people). The top 5 themes associated with low drug-induced mortality included detoxification and addiction rehabilitation services ($r=0.26$), gratitude for recovery ($r=-0.25$), thankful for treatment ($r=-0.32$), caring staff and amazing experience ($r=-0.23$), and individualized recovery programs ($r=-0.20$). The top 5 themes associated with high mortality were care from doctors or providers ($r=0.24$), rude and insensitive care ($r=0.23$), medication and prescriptions ($r=0.22$), front desk and reception experience ($r=0.22$), and dissatisfaction with communication ($r=0.21$). In the multilevel linear model, a state with a 10 deaths per 100,000 people increase in mortality was associated with a 0.30 lower average Yelp rating ($P=.005$).

Conclusions: Lower online ratings of SUD treatment facilities were associated with higher drug-induced mortality at the state level. Elements of patient experience may be associated with state-level mortality. Identified themes from online, organically derived patient content can inform efforts to improve high-quality and patient-centered SUD care.

(*JMIR AI 2023;2:e46317*) doi:[10.2196/46317](https://doi.org/10.2196/46317)

KEYWORDS

opioid use disorder; online reviews; drug-induced mortality; addiction; substance use disorder treatment; substance use disorder; patient-centered care; digital health; treatment; substance use; online review; drug use; mortality; database; addiction; detoxification; rehabilitation; communication; patient-centered

Introduction

Drug-induced mortality across the United States has continued to rise [1] from 6.2 to 21.6 age-adjusted deaths per 100,000 people over the last 20 years [2]. Recently, the Centers for Disease Control and Prevention (CDC) reported 70,630 drug overdose deaths in the United States—an average of 193 deaths every day [2]. People with substance use disorder (SUD) have higher prevalence rates of major medical conditions and a higher disease burden compared with the general population [3]. SUD-related morbidity and mortality are projected to increase over the next year [4]. There is an increased focus on ensuring that efforts to address and reduce drug-induced morbidity and mortality are patient centered to increase adoption [5,6].

To date, there are limited measures to evaluate patient preferences and priorities regarding SUD treatment [7,8], and many patients do not have access to evidence-based treatment options [9]. Patients and their families seeking SUD treatment may begin their search for an SUD treatment facility online, where they can find information about individual facilities, as well as a summary of patient-generated online reviews via popular platforms such as Google or Yelp [5]. While online reviews are not validated measures of quality of care as compared with Press Ganey or the Hospital Consumer Assessment of Healthcare Providers and Systems (HCAHPS), the use of online ratings of health care experiences continues to grow, reflecting the general trend of how consumers are seeking health-related information [10]. Prior studies of many medical settings, including essential health care facilities [11], mental health treatment facilities [12], hospitals [10], emergency departments [13], urgent care centers [13], and skilled nursing facilities, have demonstrated that online reviews may capture aspects of the patient experience that are associated with positive or negative ratings, as well as quality of care [14].

Online reviews of SUD treatment facilities may reflect information about factors associated with positive or negative patient satisfaction [15,16]. This content may provide insights to inform the development of SUD treatment performance metrics and patient-driven priorities. Evaluating this is important as understanding patient experiences is key to moving toward more patient-centered care and improved treatment services [17,18]. We sought to evaluate publicly available online reviews of US SUD treatment facilities to examine the association between online ratings of SUD treatment facilities and drug-induced mortality across the United States. We also aimed to explore if quality of care differences were reflected in reviews' narrative content. We examined the association

between thematic content of patient-generated online reviews associated with 1-star (lowest) versus 5-star (highest) ratings and drug-induced mortality.

Methods

Sample

All online reviews and ratings published on Yelp for outpatient SUD treatment facilities within the United States during September 2005 to October 2021 were collected. Facilities designated as non-SUD health facilities (eg, optometrists or retirement homes) were excluded ([Multimedia Appendix 1](#)). Consistent with prior studies on online reviews, analysis using natural language processes was used in SUD treatment facilities with 5 or more reviews [15].

We matched the list of US SUD treatment facilities to their corresponding facilities in the 2016 National Directory of Drug and Alcohol Abuse Treatment Facility Record published by the Substance Abuse and Mental Health Services Administration (SAMHSA). Matching was done using facility name and address to calculate the shortest string matching Levenshtein distance [19]. If an SUD treatment facility was not listed within the SAMHSA directory, then it was not included in the analysis.

Drug-induced mortality rates for each state were collected from the CDC WONDER Database from 2006 to 2019, and state averages were determined. Descriptive statistics were used to determine the univariate and bivariate distributions of Yelp review ratings and drug-induced mortality rates. Drug-induced mortality was treated as a continuous variable. States were considered to have “high” drug-induced mortality if their average drug-induced mortality rate was above the mean for all states. Likewise, states were considered to have “low” drug-induced mortality if their average drug-induced mortality rate was below the mean for all states.

Generating Topics, Identifying Themes, and Examining Associations With Facility Online Ratings

Latent Dirichlet allocation (LDA) is a machine learning approach that groups co-occurring words into topics. These topics are then hand-coded to identify associated themes [20]. LDA uses an unsupervised dimension reduction procedure [20] to identify latent topics among large quantities of text. The distribution of LDA topics was extracted for each facility. Themes were categorized by an independent review by 2 members of the research team (AKA and MPA), and differences were reconciled by a third member (RMM).

Ordinary least-squares regressions were performed to generate topics associated with the drug-induced mortality rates of each facility's state average. Pearson r was used to calculate effect size. For each topic generated, 10 reviews were identified. Specifically, the probability of all topics for each review was calculated, and subsequently, reviews that had the highest probability for each topic were identified. These 10 reviews were used by 3 coders (AKA, RMM, and MPA) to assign each topic a theme. The Benjamini-Hochberg P correction and $P < .05$ were used to identify significant correlations. Paired 2-tailed t tests with the Benjamini-Hochberg P correction were used to measure statistically significant associations between themes and state-level drug-induced mortality rates.

Multilevel Modeling of the Association Between Yelp Ratings and Drug-Induced Mortality Rates

Because the multilevel mixed-effects linear regression model accounts for variation at the facility level, all states with facilities with at least 5 online ratings were included ($n=51$ states). We used null random-intercepts models to calculate intraclass correlations and variance partitioning coefficients to determine the degree of clustering in ratings at the facility and state levels. The average correlation of ratings in the same state (ie, intraclass correlation) was 0.03, while that among ratings from the same facility was 0.21. Variance components analysis showed that 2% of the variance in ratings was explained at the state level, 17% was explained at the facility level, and the remaining 81% was within facilities.

Likelihood ratio tests revealed that models that accounted for clustering at both the facility and state level fit the data better than those that accounted for only the former ($\chi^2_1=137.7$, $P < .001$), only the latter ($\chi^2_1=946.5$, $P < .001$), or neither ($\chi^2_1=1405.1$, $P < .001$). Neither of the models that allowed the relationship between drug mortality and rating to vary at the state or facility level converged, so we proceeded with a 3-level, random-intercepts model with ratings nested in facilities and nested in states.

The 3-level, random-intercepts model used to assess the relationship between online review ratings and drug mortality rates integrated only 1 state-level predictor (drug-induced mortality rates), which was grand mean centered to improve interpretability of the intercept. As the outcome was on a 5-point Likert scale, we conducted a sensitivity analysis rerunning the model using a mixed-effects ordinal regression to see if it altered the results. There were no missing data for the predictor and outcome. All analyses were conducted in Stata (version 15; StataCorp).

Table 1. Descriptive statistics for low and high drug-induced mortality states included in natural language processing analyses ($n=47$).

Category	SAMHSA ^a facilities, n	Reviews, n	Reviews per facility, mean (SD)	Facility rating, mean (SD)
Low drug-induced mortality states ^b ($n=24$)	399	6853	16.03 (20.40)	3.06 (1.11)
High drug-induced mortality states ^b ($n=23$)	190	2744	13.24 (13.13)	2.64 (1.01)

^aSAMHSA: Substance Abuse and Mental Health Services Administration.

^bStates with 5 or more reviews were included in the natural language processing analyses.

Ethical Considerations

This study was considered exempt by the University of Pennsylvania institutional review board.

Results

Descriptive Statistics of Sample

A total of 589 SUD treatment facilities listed within the SAMHSA directory (6.5% of 9061 US SAMSHA-designated facilities) met the inclusion criteria of having at least 5 reviews ($n=9597$ reviews; $n=9597$ ratings). These facilities belonged to 47 states. Most facilities represented the West US census region ($n=316$), followed by the South ($n=130$), Midwest ($n=67$), and Northeast ($n=62$). The number of online ratings of SUD treatment facilities was the same as the number of online reviews (ie, each online review had a corresponding rating, so the sample included 9597 reviews and 9597 ratings).

Ratings for the 589 facilities had a bimodal distribution with peaks at a rating of 1 ($n=4546$) and 5 ($n=3649$) with a median (IQR) of 2 (2-4). The mean (SD) facility rating was 2.82 (1.87). Among these facilities, the mean (SD) state-level drug-induced mortality rate was 17.57 (5.30; range 7.54-35.01) age-adjusted deaths per 100,000 people. States were considered to have "higher than average" (ie, "high") or "lower than average" (ie, "low") drug-induced mortality if their average drug-induced mortality rate was above or below the average of 17.57 age-adjusted deaths per 100,000 people.

States With Low Drug-Induced Mortality Rates

A total of 24 (51%) of 47 states in our sample had a low drug-induced mortality rate (mean 13.40, SD 2.45 age-adjusted deaths per 100,000 people; see [Table 1](#) for descriptive statistics for low and high drug-induced mortality states).

[Tables 2](#) and [3](#) display themes, correlation coefficient, and example quotations for each theme from online reviews associated with high or low drug mortality rates. We identified 9 distinct themes associated with low drug mortality rates and 14 distinct themes associated with high drug mortality rates. The top 5 themes most correlated with a low mortality rate included the following: detox and addiction rehabilitation services ($r=-0.26$), gratitude for sobriety and recovery ($r=-0.25$), thankful for treatment ($r=-0.25$), caring staff and amazing experience ($r=-0.23$), and individualized recovery programs ($r=-0.20$; [Tables 2](#) and [3](#)). Review language correlated with positive or negative state-level drug mortality rates is displayed in [Multimedia Appendix 2](#).

Table 2. Themes across substance use disorder facilities most associated with states with low drug-induced mortality rates^a.

Theme	Drug mortality rates, Pearson <i>r</i> (95% CI)	Top words	Example reviews (redacted to maintain anonymity)
Detox and addiction rehabilitation services	-0.26 (-0.33 to -0.18)	Program, sober, recovery, detox, addiction, rehab, drug, alcohol, clean, living, house, drugs, new, meetings, step	<ul style="list-style-type: none"> • “[Facility name] changed my life! I learned about the disease of addiction and how to cope with life without the use of drugs or alcohol in this program. I couldn't be more grateful for my [Facility name] family and I continue to live a life free of drugs and alcohol by working on myself in a twelve step program. Thank you for helping me discover a better way of living.”
Gratitude for sobriety and recovery	-0.25 (-0.32 to -0.17)	Life, am, sober, years, house, grateful, today, addiction, hope, myself, saved, live, helped, learned, gave	<ul style="list-style-type: none"> • “Almost 3 years ago I moved into the [facility name]. I've been clean and sober ever since. The [facility name] gave me the structure and spiritual tools to learn how to live a life of meaning and how to be a contributing member of society sober! In July I'll celebrate 3 years continuous years of recovery and I owe my life to the 12 step program I work and the [facility name]. Thanks for everything [staff name].
Thankful for treatment	-0.25 (-0.32 to -0.17)	Life, thank, amazing, love, truly, helping, god, grateful, enough, helped, saved, beyond, special, heart, open	<ul style="list-style-type: none"> • “I am truly grateful for the care I received at the [facility name]...support staff, dietary, counselors, therapy as well as amazing facilitators. I am both humbled and grateful for my newfound sobriety. I truly hope I can carry this message to help others.
Caring staff and amazing experience	-0.23 (-0.30 to -0.15)	Recommend, recovery, house, amazing, great, highly, best, beautiful, anyone, food, truly, clients, detox, caring, comfortable	<ul style="list-style-type: none"> • “Amazing, clean facility. Caring staff, exceptional chefs. I highly recommend it. The detox house and residential house are extremely nice. The rooms are spacious and all amenities are provided.” • “This place is absolutely amazing”
Individualized recovery programs	-0.20 (-0.27 to -0.12)	Program, recovery, treatment, addiction, support, clients, group, programs, individual, approach, environment, highly, team, each, sobriety	<ul style="list-style-type: none"> • “My time spent at bayside marin has been life changing...[facility name] has given me the tools for successful recovery. The team is top notch - highly educated in the evolving field of recovery...The treatment is smart and individualized. They also offer a free alumni meeting one evening a week, which I attend often. It's a great aftercare resource.” • “The staff at [facility name] is so caring, knowledgeable and professional. Their philosophy and addiction recovery model is progressive and holistic, treating the whole person and helping them relearn how to “do life” sober and happy...” • “Facility Name” has a great vision for recovery. One size does not fit all. Finding a unique and individualized recovery path can mean the difference between temporarily quitting and truly experiencing life change.”
Appreciation of care team	-0.19 (-0.26 to -0.10)	Life, center, recovery, helped, best, recommend, amazing, truly, highly, team, love, grateful, saved, caring, hope	<ul style="list-style-type: none"> • “I don't have the words to express the depth of my gratitude for [facility name] and all the staff! I have an entirely different perspective on my whole life, and a clear understanding of myself, in a universal sense. I'm home, living my life in a light of love and compassion, thanks to the work we did at [facility name].” • “I am extremely grateful for the experience and treatment I received at [facility name]. All of the staff and therapists are extremely caring and knowledgeable. In particular [name]. He was extremely important and influential in my treatment there. My family and I will be forever grateful for him and the rest of the staff at [facility name]. If you know anyone struggling with addiction, I would highly recommend [facility name].”
Group therapy sessions	-0.14 (-0.22 to -0.06)	Therapy, day, group, week, groups, meetings, therapist, sessions, classes, once, etc, class, aa, meeting, each	<ul style="list-style-type: none"> • “I thought this class was going to be boring like most classes are. It was quite the opposite. [The class] was very interesting and educational... The instruction was awesome”

Theme	Drug mortality rates, Pearson r (95% CI)	Top words	Example reviews (redacted to maintain anonymity)
Clinic management	-0.12 (-0.20 to -0.04)	Clients, client, director, management, clinical, run, high, business, lack, completely, employees, poor, focus, communication, field	<ul style="list-style-type: none"> • “[Facility name] has had a complete reboot with leadership, programming, and staff since mid-2015. They have achieved the coveted joint commission accreditation, and hired all new credentialed therapists as well as highly trained behavioral health techs. This program in no way resembles the decision point of the past which is a real source of pride with the staff and clients. The biggest change clients will see is in the expanded programming and activities covering seven days per week.”
Case management and legal support	-0.10 (-0.18 to -0.02)	Case, manager, court, classes, class, managers, legal, client, jail, course, huge, dui, problems, ordered	<ul style="list-style-type: none"> • “something must be said about the love I received from this program that was above and beyond that which is the norm... the legal team (specifically Dr. [name]) rendered support with progress reports to the court throughout my legal proceedings, appeared in court for me and successfully got my 7 year prison sentence suspended with alternate sentencing to where I did no jail or prison time. I can’t begin to say how grateful I am to have her support along with the entire staff of [facility name]”.

^aSignificance was measured using a paired 2-tailed t test with the Benjamini-Hochberg P correction ($P < .05$).

Table 3. Themes across substance use disorder facilities most associated with states with high drug-induced mortality rates^a.

Theme	Drug mortality rates, Pearson <i>r</i> (95% CI)	Top words	Example reviews (redacted to maintain anonymity)
Dissatisfaction with length of stay and discharge process	0.11 (0.03-0.19)	Facility, days, stay, discharge, hours, without, given, during, social, worker, plan, once, friend, upon, case	<ul style="list-style-type: none"> “Ugh! Horrible! Rude staff, no individual therapy, ridiculous rules! And discharge planning? What discharge planning? You’re on your own there.” “Complete lack of discharge planning. My daughter was sent home with no follow up care plan and they wouldn’t even ship her belongings.”
Insurance, payments, and billing	0.11 (0.03-0.19)	Insurance, pay, money, bill, billing, paid, company, payment, charged, received, financial, charge, covered, check, card	<ul style="list-style-type: none"> “Do not come here – they are the worst clinic. The doctors are fine but they have you waiting forever, they screwed up our billing and wanted us to pay over \$1000 in bills they submitted to the wrong insurance company and then double billed them. Save your time and go to some other place.” “Misleading insurance coverage information – abundantly clear to me that they really only want private pay patients.” “Just received the bill for 6 days of nothing....\$10,000.00 'm telling you to stay away from this nasty dirty place. absolutely worthless!!” “Questionable billing practices at [facility name]. My husband received not one but two bills totaling over \$23,000.00. We discovered that neither bill had been submitted to insurance for payment prior to billing him directly for the full amounts.”
Therapy for co-occurring mental health disorders	0.12 (0.04-0.20)	Mental, therapy, therapist, depression, disorder, psychiatrist, health, eating, anxiety, diagnosis, inpatient, group, outpatient, disorders, social	<ul style="list-style-type: none"> “[Facility name] saved my life. I am very pleased with everything. I would recommend [Facility name] to anyone with eating disorders and mental health issues.” “Admitted for my eating disorder. Excellent physicians (especially Dr. [name]), therapists, nutritionists (specifically qualified for ED), nurses, and mental health workers. Always available to assist...They mostly use CBT (cognitive behavioral therapy) which is an effective method for all types of addictions/disorders...I'm glad...now I have the tools I need for my recovery...I highly recommend this treatment facility for patients with alcohol addiction, drug addiction, mental health disorders, and eating disorders.”
Mental health resources	0.14 (0.06-0.22)	Help, health, mental, need, issues, those, services, crisis, may, illness, willing, seek, substance, serious, deal	<ul style="list-style-type: none"> “Addiction and behavioral health issues are complex and serious. I have experienced ttc as a thorough and caring approach to improving the lives of people who ask for help.” “Very good with their counseling and resources for help.” “Caring group of mental health experts.” “Professional, kind, compassionate mental health services.”
Communication with nurse	0.16 (0.08-0.24)	Told, said, didn't, then, got, nurse, left, asked, came, took, down, couldn't, mom, let, saying	<ul style="list-style-type: none"> “I never once saw my nurse after being in the room for an hour. They were too busy gossiping at the nurses station so for the reviews 9 months ago that had a response from the hospital saying they'll work on it is BS they'll still prioritize talking instead of taking care of patients”
Patients feeling restrained or held against their will	0.17 (0.08-0.24)	Patients, down, leave, please, keep, hold, unit, send, prison, against, worse, police, sleep, admitted, allowed	<ul style="list-style-type: none"> “Awful awful place. do not go here. Go to your regular psych or doctor before you ever step foot in this institution. It is more like a prison than a mental health facility.” “A prison-like “health” facility where you may come out of a worse...wonder why the only place with open beds and kinda warned by medical hospital staff some patients without prison mentality or if they're not not totally insane please watch your backs especially vulnerable and young”

Theme	Drug mortality rates, Pearson r (95% CI)	Top words	Example reviews (redacted to maintain anonymity)
Patient complaints and privacy concerns	0.17 (0.09-0.25)	Patient, information, complaint, state, against, due, name, records, refused, privacy, report, unprofessional, file, director, law	<ul style="list-style-type: none"> “I recommend against using this company. I went for an assessment on my own accord and paid myself. The final report issued had many errors that they refused to correct...Giving a false assessment and not correcting it after the errors were pointed out. I recommend you go somewhere more professional.” “If I could give no stars I would. Hipaa violations, unethical, incorrect medications, unprofessional and beyond belief still in business. Buyer beware. My records were altered and I have to get them legally rectified...Copious violations.”
Communication regarding appointments and office closures	0.21 (0.13-0.28)	Told, said, called, asked, then, see, needed, next, until, pm, morning, friday, monday, today, hour	<ul style="list-style-type: none"> “It would have been nice if someone had told us you guys closed early today!!!! My husband had an appointment at 4:20pm and when we got to the clinic at 4pm, security said it was closed!!!! His appointment is not until 4:20pm. I called the call center and even they said the clinic closed at 5pm!!!! We wasted our money for parking and most importantly our time!!!!!!” “Awful experience! Zero stars if possible!!! Their intake hours are supposed to be Monday through Friday 8:30am to 2pm. I was told by the [curse word] that answered the phones that the intake appointments take 1-2 hours. I went there at 1:50pm...They said it was too late and to come back tomorrow. How could it be too late? Apparently, the intake appointments are 2-4 hours now. They have a new system. yeah, a new waste of my time system. No thanks.”
Wait time for appointments	0.21 (0.13-0.28)	Appointment, time, minutes, wait, office, waiting, appointments, hour, before, late, scheduled, schedule, long, waited, seen	<ul style="list-style-type: none"> “The [facility name] go can shove it- if you arrive 5 minutes late- they tell you to go away. If you want an appointment you're looking months out- but show up to your appointment 20 minutes early- and you'll be waiting an hour after your appointment time...” “Appointment time - 12:30pm, arrival time - 12:15pm, current time - 1:39pm and I'm still waiting!! Because this place is located in a predominantly black and hispanic neighborhood, these people think they can disrespect our time and have us waiting here for over an hour!! Stay away!” “Very poorly managed time wise. My first appointment was over two hours late from the scheduled time. huge co-pay. My second appointment was also over an hour late even though it was maybe a 10 minute consultation. My yet to be third appointment has been rescheduled twice, once 15 minutes beforehand.”
(Dissatisfaction with) phone calls and lines of communication	0.21 (0.13-0.29)	Call, phone, called, back, calls, someone, left, times, number, calling, message, answer, messages, speak, hold	<ul style="list-style-type: none"> “No one answers the phone or calls you back. I can't get a prescription refilled. The automated system is like the people...doesn't work. Don't waste your time.” “This company, after an initial consultation and intent to become a patient, did not respond to my multiple emails and several calls and voicemails for over 2 months.”
Front desk and reception experience	0.22 (0.14-0.29)	Rude, front, desk, treated, unprofessional, attitude, service, extremely, woman, worst, horrible, speak, lady, ask, name	<ul style="list-style-type: none"> “Extremely rude and unhelpful. When I called to make an appointment the therapist was dry, rude and clearly uninterested.” “The medical center is great and the staff however the front desk woman is extremely rude, cold and disrespectful. It's a shame to have someone like her representing [facility name].”
Medication choices and prescription refills	0.22 (0.14-0.30)	Medication, meds, doctor, medications, off, drug, psychiatrist, prescription, prescribed, pain, anxiety, drugs, withdrawal, med, effects	

Theme	Drug mortality rates, Pearson r (95% CI)	Top words	Example reviews (redacted to maintain anonymity)
			<ul style="list-style-type: none"> “This office is unhelpful and apathetic about refilling prescriptions. I am on week two now of daily phone calls to get a non-narcotic antidepressant prescription refilled. There is no reason why it shouldn't be filled, yet my calls remain unreturned and the soonest I can see a doctor is three weeks despite having 0 days left of my meds, which they know.” “The standard of care is dismally low. Gaslighting by doctors, patients being told to go OD so they can qualify for care, and patients being put out addicted to a cocktail of pills without informed consent regarding withdrawal effects or tapering regimens. This place exists to make money, not to heal.” “The psychiatrist hastily prescribed a narcotic that had a negative interaction with my other medication, she ignored the list of meds.”
Rude and insensitive care	0.23 (0.15-0.31)	Go, here, don't, give, worst, ever, horrible, stars, rude, anyone, star, zero, nothing, please, worse	<ul style="list-style-type: none"> “The nurses were horrible, unattentive, with no compassion whatsoever. Worst hospital experience ever. Wow! I hope I never have to go there again.” “Liars liars liars! incompetent, obnoxious apathetic, rude, arrogant. Seriously, the worst of humanity works here.”

^aSignificance was measured using a paired 2-tailed t test with the Benjamini-Hochberg P correction ($P < .05$).

States With High Drug-Induced Mortality Rates

A total of 23 (49%) of 47 states in our sample had a high drug-induced mortality rate (mean 21.92, SD 3.69 age-adjusted deaths per 100,000 people; [Table 1](#)).

The top 5 themes most correlated with high drug mortality rates included care from doctors or providers ($r=0.24$), rude and insensitive care ($r=0.23$), medication choices and prescription refills ($r=0.22$), front desk and reception experience ($r=0.22$), and (dissatisfaction with) phone calls and lines of communication ($r=0.21$; [Tables 2 and 3](#)).

Associations Between Review Ratings and Drug-Induced Mortality Rates

Across all states ($n=11$, 941 ratings), the mean (SD) mortality rate was 17.1 (5.5; range 6.8-35.0) age-adjusted deaths per 100,000 people. Multilevel modeling revealed that in a typical facility in a state with an average drug mortality rate, the predicted average Yelp rating was 2.6 (95% CI 2.5-2.8) out of 5. On average, there was a negative association between drug mortality rate and Yelp ratings ($b=-0.03$, 95% CI -0.05 to -0.01 ; $P=.005$). Therefore, a state with a 10 deaths per 100,000 people increase in drug-induced mortality was associated with a 0.30 points lower average Yelp rating. This negative association was replicated in the mixed-effects ordinal regression model ($b=-0.04$, 95% CI -0.07 to -0.01 , $P=.004$).

Discussion

Principal Findings

This study analyzed the association between online ratings and narrative review content from online reviews of US SUD treatment facilities and drug-induced mortality data from the CDC. The study has 2 main findings. First, we found that the average negative online ratings of SUD treatment facilities were

associated with higher drug-induced mortality. Second, there were marked differences in the themes expressed between high versus low mortality states. These findings provide insights about the gap that persists in understanding the associations between online reviews and drug-induced mortality outcomes. Further, these results may help amplify patient-generated perceptions of poor quality of SUD care that may contribute to increased drug-induced mortality.

For every 10 deaths per capita increase in drug-induced mortality, the Yelp rating is expected to be 0.3 points lower. This is important, as little research has been conducted to closely examine the association between the online ratings and morbidity and mortality outcomes in the context of SUD treatment [11]. Consistent with a prior report that found that higher online ratings of essential health care facilities were associated with lower mortality [11], our findings suggest that online ratings may serve as a proxy for some components of quality of care such as communication with patients or availability of evidence-based treatments. This work also provides evidence that tools such as ATLAS [21], a website developed to help patients find and compare SUD treatment facilities, may have value in guiding patients to care options that fit their needs and preferences.

Recently, the Shatterproof foundation developed National Principles of Care for addiction treatment, evidence-based practices to improve outcomes for individuals with SUD [22]. Themes associated with low mortality were consistent with these principles. For example, their second principle, “A personal plan for every patient,” matched the theme “individualized recovery programs.” This theme is also in line with a recent partnership between Shatterproof, the American Society of Addiction Medicine, and OpenBeds to create a free, 13-item assessment to determine what type of SUD treatment aligns best with each patient's needs [23].

These findings provide insights into aspects of patient experience within SUD care that are often difficult to capture with numerical surveys including a focus on “caring staff” and “communication.” Themes associated with high mortality states often pertained to poor communication and low-quality or non-evidenced-based care. Many of these identified themes can guide areas of improvement regarding the delivery of patient-centered and high-quality care. The identified themes indicate aspects of the patient experience that may contribute to high and low state-level mortality. Ultimately, these results underscore a process to unify patients’ “digital voices” to improve and inform treatment for SUD.

Limitations

This study has several limitations. Reviews in the sample represent a small proportion of a facility’s patients, and facilities included represent a very small proportion of the SUD treatment infrastructure. Further, online reviews may not be representative of the population seen at each facility because Yelp does not verify the identity of the user posting a rating or review. Therefore, the use of only Yelp reviews as a source of online ratings and reviews may limit the impact of our findings. Additionally, 4 states (including Washington DC) did not have SUD treatment facilities with more than 5 reviews, limiting conclusions that can be drawn about the association between themes in online ratings and mortality in those states. Further, consistent with previously published methods [10-13,15] to analyze thematic online review content, the analyses in this study were not stratified by year, which limits conclusions that can be drawn. Specifically, our data are limited by the fact that the distribution of ratings by year is slightly skewed toward later years when reviews of health centers on Yelp became more popular. Other limitations of this study include its retrospective design, selection bias, and responder bias. A final limitation is that due to our sample size, our analyses were limited to mortality data at the state level despite the fact that county-level mortality data are generally available, so we could not explore facility-level services or practices that may contribute to high drug mortality. If more reviews become available, a county-level analysis in the future may provide more granular results. Our

team attempted to run a similar analysis at the county level, but the intersection of mortality data from CDC and review data from Yelp was very small. Likewise, there may be possible heterogeneity across SUD populations in different states that limits the impact of these findings, as well as differences in state-level investment in SUD care and responses to drug-induced mortality rates that vary depending on state-level priorities and budgetary restrictions. Although state policy likely is linked to mortality, state-level policy differences were not likely captured in the patient-generated online content analyzed in this study.

This study also has strengths. Online review platforms serve as an organic, democratizing, and accessible space for patients to document their care experiences with rich narratives. While reviews are not representative, Yelp uses software in place to filter out inappropriate or inaccurate reviews. Moreover, the anonymity of reviews may encourage patients to express the true realities of their experiences without fear that it will impact their care. Therefore, analyses of online review content can provide insights to improve patient experiences and treatment delivery that may not be captured by numerical surveys or patient experiences surveys where patients may be concerned that their anonymity is not protected.

Conclusions

At the state level, mean negative online ratings of SUD treatment facilities were associated with higher drug-induced mortality. Additionally, unique narrative content themes were identified online reviews across states with low or high mortality. Online reviews of SUD treatment facilities provide an opportunity to investigate and understand elements of the patient experience, quality of care, and state level mortality. The themes generated from online, organically derived patient content can inform and improve patient-centered care for SUD treatment. Future efforts to integrate these themes into the development of an SUD treatment facility-based performance and quality measures for SUD treatment may help to further elucidate what aspects of patient care may promote or improve both patient satisfaction and drug-induced mortality.

Acknowledgments

The authors would like to acknowledge the faculty and staff of the Center for Digital Health and the Center for Emergency Care Policy and Research at the University of Pennsylvania for their support of this work. Funding was provided by National Institutes of Health, National Institute on Drug Abuse (NIH NIDA; 1R21DA050761). The authors would also like to thank Nina Sokolovic for her guidance regarding multilevel modeling and overall support of the lead author’s research initiatives.

Data Availability

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

Authors' Contributions

APP and SCG lead the natural language processing analyses and provided guidance on the statistical analyses led by MPA. RMM, ZFM, and AKA lead the study design and provided guidance to MPA, SCG, and APP about the analyses. Themes were categorized by independent review by 2 members of the research team (AKA and MPA) and differences reconciled by a third (RMM). All authors wrote parts of the article and provided revisions to this manuscript. All authors read and approved the final manuscript.

Conflicts of Interest

RMM is currently supported as principal investigator by the National Institutes of Health (NIH) National Institute on Drug Abuse (NIDA; award 1R21DA050761); NIH National Heart, Lung and Blood Institutes of Health (awards K24-HL157621 and R01HL14184401); and the National Institute of Mental Health (award R01MH127686). None of the other authors have competing interests to declare.

Multimedia Appendix 1

Excluded facilities based on Yelp category label.

[[DOCX File, 26 KB - ai_v2i1e46317_app1.docx](#)]

Multimedia Appendix 2

Words most associated with online reviews in states with (A) high and (B) low drug-induced mortality rates. Relative font size represents stronger correlation with high or low mortality. Increased frequency of word use is represented by darker shading.

[[PNG File, 219 KB - ai_v2i1e46317_app2.png](#)]

References

1. Substance abuse and addiction statistics. National Center for Drug Abuse Statistics. 2022. URL: <https://drugabusestatistics.org/> [accessed 2022-03-21]
2. Multiple cause of death 1999-2019 on CDC WONDER online database, released in 2020. Data are from the multiple cause of death files, 1999-2019, as compiled from data provided by the 57 vital statistics jurisdictions through the Vital Statistics Cooperative Program. Centers for Disease Control and Prevention, National Center for Health Statistics. 2021. URL: <https://wonder.cdc.gov/mcd-icd10.html> [accessed 2023-11-07]
3. Bahorik AL, Satre DD, Kline-Simon AH, Weisner CM, Campbell CI. Alcohol, cannabis, and opioid use disorders, and disease burden in an integrated health care system. *J Addict Med* 2017;11(1):3-9 [FREE Full text] [doi: [10.1097/ADM.0000000000000260](https://doi.org/10.1097/ADM.0000000000000260)] [Medline: [27610582](https://pubmed.ncbi.nlm.nih.gov/27610582/)]
4. Provisional drug overdose death counts. Centers for Disease Control and Prevention. URL: <https://www.cdc.gov/nchs/nvss/vsrr/drug-overdose-data.htm> [accessed 2022-03-21]
5. Agarwal AK, Guntuku SC, Meisel ZF, Pelullo A, Kinkle B, Merchant RM. Analyzing online reviews of substance use disorder treatment facilities in the USA using machine learning. *J Gen Intern Med* 2022;37(4):977-980 [FREE Full text] [doi: [10.1007/s11606-021-06618-7](https://doi.org/10.1007/s11606-021-06618-7)] [Medline: [33728567](https://pubmed.ncbi.nlm.nih.gov/33728567/)]
6. Marchand K, Beaumont S, Westfall J, MacDonald S, Harrison S, Marsh DC, et al. Conceptualizing patient-centered care for substance use disorder treatment: findings from a systematic scoping review. *Subst Abuse Treat Prev Policy* 2019;14(1):37 [FREE Full text] [doi: [10.1186/s13011-019-0227-0](https://doi.org/10.1186/s13011-019-0227-0)] [Medline: [31511016](https://pubmed.ncbi.nlm.nih.gov/31511016/)]
7. Garnick DW, Horgan CM, Acevedo A, McCorry F, Weisner C. Performance measures for substance use disorders—what research is needed? *Addict Sci Clin Pract* 2012;7(1):18 [FREE Full text] [doi: [10.1186/1940-0640-7-18](https://doi.org/10.1186/1940-0640-7-18)] [Medline: [23186374](https://pubmed.ncbi.nlm.nih.gov/23186374/)]
8. Weisner C, Campbell CI, Altschuler A, Yarborough BJH, Lapham GT, Binswanger IA, et al. Factors associated with Healthcare Effectiveness Data and Information Set (HEDIS) alcohol and other drug measure performance in 2014-2015. *Subst Abuse* 2019;40(3):318-327 [FREE Full text] [doi: [10.1080/08897077.2018.1545728](https://doi.org/10.1080/08897077.2018.1545728)] [Medline: [30676915](https://pubmed.ncbi.nlm.nih.gov/30676915/)]
9. County buprenorphine access in the United States. Shatterproof. URL: <https://www.shatterproof.org/our-work/advocacy/research-reports/buprenorphine-access> [accessed 2022-03-21]
10. Ranard BL, Werner RM, Antanavicius T, Schwartz HA, Smith RJ, Meisel ZF, et al. Yelp reviews of hospital care can supplement and inform traditional surveys of the patient experience of care. *Health Aff (Millwood)* 2016;35(4):697-705 [FREE Full text] [doi: [10.1377/hlthaff.2015.1030](https://doi.org/10.1377/hlthaff.2015.1030)] [Medline: [27044971](https://pubmed.ncbi.nlm.nih.gov/27044971/)]
11. Stokes DC, Pelullo AP, Mitra N, Meisel ZF, South EC, Asch DA, et al. Association between crowdsourced health care facility ratings and mortality in US counties. *JAMA Netw Open* 2021;4(10):e2127799 [FREE Full text] [doi: [10.1001/jamanetworkopen.2021.27799](https://doi.org/10.1001/jamanetworkopen.2021.27799)] [Medline: [34665240](https://pubmed.ncbi.nlm.nih.gov/34665240/)]
12. Stokes DC, Kishton R, McCalpin HJ, Pelullo AP, Meisel ZF, Beidas RS, et al. Online reviews of mental health treatment facilities: narrative themes associated with positive and negative ratings. *Psychiatr Serv* 2021;72(7):776-783 [FREE Full text] [doi: [10.1176/appi.ps.202000267](https://doi.org/10.1176/appi.ps.202000267)] [Medline: [34015944](https://pubmed.ncbi.nlm.nih.gov/34015944/)]
13. Agarwal AK, Mahoney K, Lanza AL, Klinger EV, Asch DA, Fausti N, et al. Online ratings of the patient experience: emergency departments versus urgent care centers. *Ann Emerg Med* 2019;73(6):631-638 [FREE Full text] [doi: [10.1016/j.annemergmed.2018.09.029](https://doi.org/10.1016/j.annemergmed.2018.09.029)] [Medline: [30392737](https://pubmed.ncbi.nlm.nih.gov/30392737/)]
14. Ryskina KL, Andy AU, Manges KA, Foley KA, Werner RM, Merchant RM. Association of online consumer reviews of skilled nursing facilities with patient rehospitalization rates. *JAMA Netw Open* 2020;3(5):e204682 [FREE Full text] [doi: [10.1001/jamanetworkopen.2020.4682](https://doi.org/10.1001/jamanetworkopen.2020.4682)] [Medline: [32407501](https://pubmed.ncbi.nlm.nih.gov/32407501/)]
15. Agarwal AK, Wong V, Pelullo AM, Guntuku S, Polsky D, Asch DA, et al. Online reviews of specialized drug treatment facilities-identifying potential drivers of high and low patient satisfaction. *J Gen Intern Med* 2020;35(6):1647-1653 [FREE Full text] [doi: [10.1007/s11606-019-05548-9](https://doi.org/10.1007/s11606-019-05548-9)] [Medline: [31755009](https://pubmed.ncbi.nlm.nih.gov/31755009/)]

16. Merchant RM, Volpp KG, Asch DA. Learning by listening-improving health care in the era of Yelp. *JAMA* 2016 Dec 20;316(23):2483-2484 [FREE Full text] [doi: [10.1001/jama.2016.16754](https://doi.org/10.1001/jama.2016.16754)] [Medline: [27997663](https://pubmed.ncbi.nlm.nih.gov/27997663/)]
17. Korthuis PT, Gregg J, Rogers WE, McCarty D, Nicolaidis C, Boverman J. Patients' reasons for choosing office-based buprenorphine: preference for patient-centered care. *J Addict Med* 2010;4(4):204-210 [FREE Full text] [doi: [10.1097/ADM.0b013e3181cc9610](https://doi.org/10.1097/ADM.0b013e3181cc9610)] [Medline: [21170143](https://pubmed.ncbi.nlm.nih.gov/21170143/)]
18. Mark TL, Hinde J, Henretty K, Padwa H, Treiman K. How patient centered are addiction treatment intake processes? *J Addict Med* 2021;15(2):134-142 [FREE Full text] [doi: [10.1097/ADM.0000000000000714](https://doi.org/10.1097/ADM.0000000000000714)] [Medline: [32826618](https://pubmed.ncbi.nlm.nih.gov/32826618/)]
19. Heeringa WJ. Measuring Dialect Pronunciation Differences Using Levenshtein Distance. Groningen: University Library Groningen; 2004.
20. Blei D, Ng A, Jordan M. Latent Dirichlet allocation. *Advances in Neural Information Processing Systems 14 (NIPS 2001)*. 2001. URL: <https://proceedings.neurips.cc/paper/2001/file/296472c9542ad4d4788d543508116cbc-Paper.pdf> [accessed 2023-11-07]
21. Shatterproof Treatment Atlas. URL: <https://www.treatmentatlas.org/> [accessed 2022-03-21]
22. Shatterproof national principles of care. Shatterproof. URL: <https://www.shatterproof.org/shatterproof-national-principles-care> [accessed 2022-03-21]
23. Get a treatment recommendation. Shatterproof. URL: <https://www.shatterproof.org/find-help/determine-treatment-needs> [accessed 2022-03-19]

Abbreviations

CDC: Centers for Disease Control and Prevention

HCAHPS: Hospital Consumer Assessment of Healthcare Providers and Systems

LDA: latent Dirichlet allocation

SAMHSA: Substance Abuse and Mental Health Services Administration

SUD: substance use disorder

Edited by G Luo; submitted 06.02.23; peer-reviewed by Q Dong, S Zeng; comments to author 16.05.23; revised version received 29.09.23; accepted 02.10.23; published 29.12.23.

Please cite as:

Abrams MP, Merchant RM, Meisel ZF, Pelullo AP, Chandra Guntuku S, Agarwal AK

Association Between Online Reviews of Substance Use Disorder Treatment Facilities and Drug-Induced Mortality Rates: Cross-Sectional Analysis

JMIR AI 2023;2:e46317

URL: <https://ai.jmir.org/2023/1/e46317>

doi: [10.2196/46317](https://doi.org/10.2196/46317)

PMID: [38875553](https://pubmed.ncbi.nlm.nih.gov/38875553/)

©Matthew P Abrams, Raina M Merchant, Zachary F Meisel, Arthur P Pelullo, Sharath Chandra Guntuku, Anish K Agarwal. Originally published in JMIR AI (<https://ai.jmir.org>), 29.12.2023. This is an open-access article distributed under the terms of the Creative Commons Attribution License (<https://creativecommons.org/licenses/by/4.0/>), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR AI, is properly cited. The complete bibliographic information, a link to the original publication on <https://www.ai.jmir.org/>, as well as this copyright and license information must be included.

Review

The Application of Artificial Intelligence in Health Care Resource Allocation Before and During the COVID-19 Pandemic: Scoping Review

Hao Wu¹, MA; Xiaoyu Lu², DPhil; Hanyu Wang², BA, BEc, MPhil

¹Department of Politics and International Relations, University of Oxford, Oxford, United Kingdom

²School of International Studies, Peking University, Beijing, China

Corresponding Author:

Hanyu Wang, BA, BEc, MPhil

School of International Studies

Peking University

No 5 Yiheyuan Road

Haidian District

Beijing, 100871

China

Phone: 86 13261712766

Email: wang.hanyu@outlook.com

Abstract

Background: Imbalanced health care resource distribution has been central to unequal health outcomes and political tension around the world. Artificial intelligence (AI) has emerged as a promising tool for facilitating resource distribution, especially during emergencies. However, no comprehensive review exists on the use and ethics of AI in health care resource distribution.

Objective: This study aims to conduct a scoping review of the application of AI in health care resource distribution, and explore the ethical and political issues in such situations.

Methods: A scoping review was conducted following the PRISMA-ScR (Preferred Reporting Items for Systematic Reviews and Meta-Analyses extension for Scoping Reviews). A comprehensive search of relevant literature was conducted in MEDLINE (Ovid), PubMed, Web of Science, and Embase from inception to February 2022. The review included qualitative and quantitative studies investigating the application of AI in health care resource allocation.

Results: The review involved 22 articles, including 9 on model development and 13 on theoretical discussions, qualitative studies, or review studies. Of the 9 on model development and validation, 5 were conducted in emerging economies, 3 in developed countries, and 1 in a global context. In terms of content, 4 focused on resource distribution at the health system level and 5 focused on resource allocation at the hospital level. Of the 13 qualitative studies, 8 were discussions on the COVID-19 pandemic and the rest were on hospital resources, outbreaks, screening, human resources, and digitalization.

Conclusions: This scoping review synthesized evidence on AI in health resource distribution, focusing on the COVID-19 pandemic. The results suggest that the application of AI has the potential to improve efficacy in resource distribution, especially during emergencies. Efficient data sharing and collecting structures are needed to make reliable and evidence-based decisions. Health inequality, distributive justice, and transparency must be considered when deploying AI models in real-world situations.

(JMIR AI 2023;2:e38397) doi:[10.2196/38397](https://doi.org/10.2196/38397)

KEYWORDS

artificial intelligence; resource distribution; health care; COVID-19; health equality; eHealth; digital health

Introduction

Global responses to COVID-19 are converging with the use of digital health and algorithms based on artificial intelligence (AI), impacting health care systems around the world [1]. AI

was partially founded by Alan Turing, and a machine or a process that could demonstrate intelligent behaviors in cognitive tasks, which can pass the Turing test, would be deemed as AI [2]. Multiple AI techniques, such as fuzzy expert systems and Bayesian networks, have been applied both virtually and physically in the health care field [3]. For example, clinical

pathway analysis, a critical area in ensuring standard medical procedures, can be analyzed by pattern-mining procedures [4]. Resource distribution includes the distribution of resources at strategic, tactical, and operational levels and is a key issue in health policy [5,6].

Luengo-Oroz et al proposed that the application of AI during the COVID-19 pandemic can be broken down into 3 scales: molecular, clinical, and societal [7]. At the molecular level, protein structure prediction, novel nucleic acid testing, drug repurposing, and drug discovery all rely on AI and deep-learning algorithms [7-9]. At the clinical level, diagnosis, treatment, and prognosis all benefit from AI. For example, AI-based computed tomography diagnosis has been widely applied for identifying COVID cases [7,10,11], alongside robotics and telemedicine that facilitate clinical processes. At the societal level, AI is applied in epidemiological research and social policymaking. In particular, AI-based case forecasting has been in use since the beginning of the pandemic [7,12]. The application of AI at the societal level can stratify population risk, facilitate diagnosis and testing, support the design of trials and drugs, and inform policymaking, relieving the burden of COVID-19 on health care systems and helping the society to better respond to the pandemic [1].

The application of AI to decision-making processes in health care systems significantly precedes the COVID-19 pandemic [7,13]. Health policy aims at providing health care to the population, and the decision-making process aims to address 2 core issues: screening and diagnosis, and treatment and monitoring [7]. These 2 tasks are essential to the entire health care system. The policymaking process includes hypothesis generation, hypothesis testing, and action (or policy). AI can learn from past data, including health records, past insurance claims, and disease incidence and prevalence, to improve hypothesis generation and testing, and thus improve the quality of health care policymaking [7].

In the health care system, resource distribution is an essential issue for policymakers, as resources are always scarce [14]. For example, Kong et al argued that the primary problem in China's health care system is the lack of high-quality health resources and the consequent supply-demand imbalance. They maintain that AI could benefit from China's enormous data and has the potential to improve this unequal distribution of health resources [14].

During the COVID-19 pandemic, imbalanced health care resource distribution has been one of the central issues causing unequal health outcomes and political tension [15,16]. Ji et al observed that the higher COVID case-fatality rate in Wuhan city and Hubei province compared with other parts of China at the beginning of the pandemic could potentially be attributed to health care resource scarcity [16]. Edejer et al projected that the cost of health care resources to combat the pandemic would continue to rise in low- and middle-income countries, and concluded that a comprehensive system of resource distribution is necessary [15].

Health care resource distribution is determined by the supply-demand relationship, logistics, and governance structure

[17,18]. Using the COVID-19 response as an example, the severity of the pandemic can determine the health care resources required in each location, but the resources might not be distributed according to need [18]. AI can be applied to study supply-demand, logistics, and patient characteristics, but the ethics and implications of the use of AI in policymaking remain important issues [7].

Currently, there are no comprehensive reviews to provide an overall picture of the literature on the application of AI in resource distribution in health care settings, particularly with regard to societal and ethical aspects. This study aims to conduct a scoping review on the application of AI in health care resource distribution, particularly during the COVID-19 pandemic and to explore the ethics and implications of AI in health policymaking with regard to resource distribution.

Methods

Scoping Review Design

This scoping review follows the framework proposed by Arksey and O'Malley [19]. Briefly, the review has the following 5 stages: (1) identifying the research question, "What are the roles of AI and machine learning in the allocation of health care resources, before and during the COVID-19 pandemic?"; (2) identifying suitable studies; (3) selecting studies for review; (4) consolidating the data; and (5) summarizing and reporting the results. This study complies with the Preferred Reporting Items for Systematic Reviews and Meta-Analyses extension for Scoping Reviews (PRISMA-ScR) [20] for reporting scoping review results.

Data Source and Search Strategy

Searches were conducted in MEDLINE (Ovid), PubMed, Web of Science, and Embase from inception to February 2022. The search featured 2 key terms: (1) *artificial intelligence*, including related terms such as *big data* and *algorithm*, and (2) *health care resource allocation*. The search terms were used with the "explode" feature where applicable. For example, in MEDLINE and Embase, we used *exp artificial intelligence/* and *exp resource allocation/*, and in PubMed, relevant MeSH (Medical Subject Heading) terms were used. The search was individually designed and adapted for each database.

Study Selection

Inclusion and exclusion criteria were defined a priori. This scoping review includes qualitative and quantitative studies investigating the application of AI in health care resource allocation. Studies that are not relevant to AI or health care resource allocation were excluded, as were duplicate studies. The inclusion and exclusion criteria are summarized in [Table 1](#).

Selection was conducted in 2 steps. First, titles and abstracts were screened for topic relevance and study design. Second, full texts of the remaining studies were screened to check for eligibility. All of the study selection processes were conducted in EndNote X9 (Clarivate).

Table 1. Inclusion and exclusion criteria.

Criterion	Inclusion	Exclusion
Type of study	Qualitative, quantitative, mixed method, and review studies in peer-reviewed journals	Letters, comments, conference abstracts, editorials, and theses
Language	English	All other languages
Study variables	Includes (1) artificial intelligence/machine learning and relevant terms and (2) allocation of health care resources	Does not include (1) artificial intelligence/machine learning and relevant terms or (2) allocation of health care resources
Study context	Health care resource allocation at either the population level or hospital level	All other resource allocation scenarios

Data Consolidation

Selected studies were input into NVivo 12 (QSR International) for labeling and coding. Authors coded data of interest from the articles in NVivo 12 and extracted information regarding study author, study design, location, context, aim, main result, AI method under study, resource allocation situation, and policymaking relevance into a standardized Excel (Microsoft Corp) form.

Summarizing the Results

We employed an inductive approach to summarize the results from the included studies. First, the selected papers were grouped into 2 types: (1) studies of model development and validation of AI-based algorithms applied to health care resource distribution, and (2) qualitative studies, theoretical discussions, and review studies of the application of AI in health care resource distribution. For studies of model development and validation, we extracted the study objectives, resource distribution situations, AI model input variables, and policy relevance. For studies in the second category, objectives, resource distribution situations, discussed topics, and policy relevance were extracted. We further divided the input variables

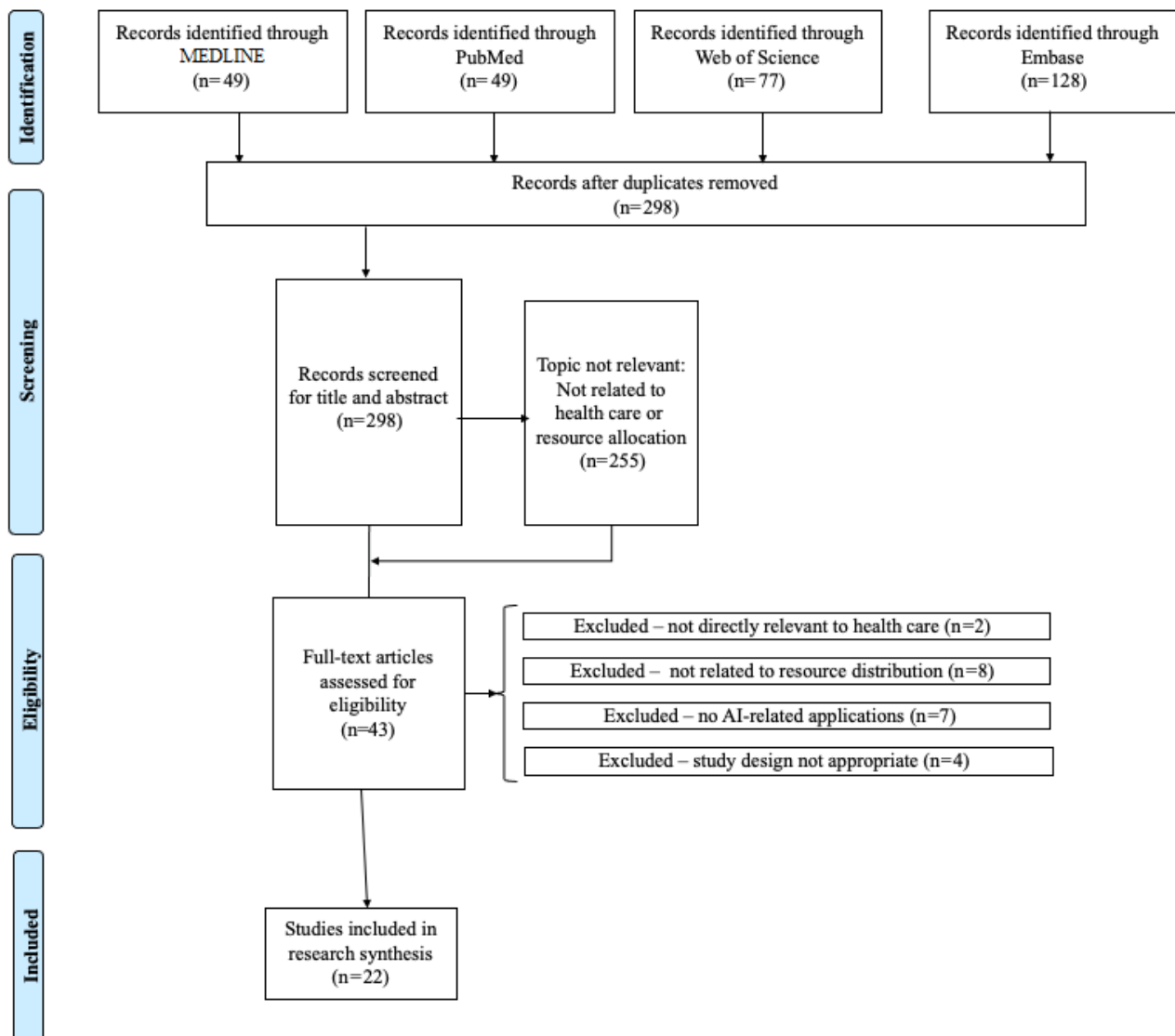
of the studies of model development and validation into 2 predefined categories: (1) ecological variables or variables at the group level, which included variables depicting characteristics at the population level, such as infant mortality in a region, local economic development, or disease prevalence and incidence; and (2) individual variables, which included variables that define individual characteristics such as diagnosis and age.

Results

Selected Studies

In total, 298 studies were identified in 4 databases after removing duplicates. After 1 round of screening for titles and abstracts, 255 studies were excluded due to irrelevant topics and unsuitable study designs. This left 43 studies for full-text screening. Of these, 2 were excluded because they were not directly relevant to health care, 8 because they were not related to resource distribution, 7 because they did not feature applications of AI, and 4 because of an inappropriate study design. In the end, 22 studies remained for qualitative synthesis. The PRISMA flow diagram for study selection is presented in [Figure 1](#).

Figure 1. PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) flow diagram of the study. AI: artificial intelligence.



Summary of the Characteristics of Studies on Model Development

The characteristics of the included studies on model development are summarized in Table 2. The included studies were published between 2013 and 2021. Of the 22 included studies, 9 focused on model development and validation [21-30]. Of these, 5 studies were conducted in emerging economies, including 2 in China [27,29], 2 in Brazil [25,28], and 1 in Ecuador [26]. In developed countries, 3 studies were conducted. These included 1 in Germany [23], 1 in the United Kingdom [22], and 1 in the United States with a validation data set in China [24]. One study was applied to a global context [21].

Of the 9 studies, 4 focused on resource distribution at the health system level, including financial resources for public health in Brazil [25], health care resource distribution in health planning in Ecuador [26], medical resource allocation in the hierarchical health system in China [29], and medical equipment allocation in the global COVID-19 pandemic [21]. The remaining 5 studies focused on resource allocation at the hospital level, including bed allocation in a London hospital [22], day resources and bed allocation in a hospital in Munich, Germany [23], human resources and medical materials in a public hospital in China [27], medical resource allocation in a hospital in the capital of State of Minas Gerais in Brazil [28], and medical resource allocation in clinics for COVID-19 patients in New York [24].

Table 2. Characteristics of the included studies on model development and validation.

Reference	Objectives	Resource allocation situation	Input variables
Rosas et al (2013) [25]	To construct a financial resource allocation model using an artificial neural network	Financial resources for public health in Brazil	Mortality characteristics, proportion of teenage mothers, proportion of inadequate prenatal care, fertility rate, Gini index, proportion of elderly people in the population, literacy rate, financing capacity per capita, percentage of people with income below half minimum wage, percentage of urban households with basic sanitation, and proportion of urban households served by garbage collection
Belciug & Gorunescu (2015) [22]	To propose a bed allocation and financial resource utilization strategy through queuing modeling and evolutionary computation	Bed allocation and financial resource utilization in the geriatric department of a London hospital	Bed inventory, arrival rate, mean service time, patient flow parameters, and holding and penalty cost and other cost considerations
Gartner & Padman (2015) [23]	To evaluate how early determination of diagnosis-related groups can be used for better allocation of scarce hospital resources	Hospital resources, including day resources and overnight resources (beds), validated in a mid-sized hospital near Munich, Germany	Primary and secondary diagnoses, clinical procedures, age, gender, and weight in newborns
Velez et al (2016) [26]	To present an artificial intelligence-based health planning model based on data from geospatial systems	Health care resource distribution in health planning in Ecuador	Geospatial variables based on the social determinants of health and geospatial patterns of territorial distribution in the allocation of equipment, supplies, and health services in relation to the availability, accessibility, and need of the population
Xu et al (2018) [27]	To propose a health resource allocation model based on mass customization to maximize revenue and customization	Allocation of doctors and other medical resources in a public hospital system in China	Distribution of medical stations, professional level of doctors (salary and seniority), patient preferences and illness severity, medical cost, and revenue
Yousefi et al (2018) [28]	To present a model based on agent-based simulation, machine learning, and a genetic algorithm for allocation of medical resources in emergency departments	Medical resource allocation in a teaching hospital in the capital of State of Minas Gerais in Brazil	Number of receptionists in the reception area; number of triage nurses in the triage room; number of laboratory technicians in the laboratory and X-ray room; and number of doctors, nurses, and nurse technicians in the suturing yellow zone, orthopedics department, surgical department, and clinical emergency area.
Zhang et al (2018) [29]	To propose a framework introducing a novel approach to multi-attribute decision-making problems in the picture fuzzy context	Medical resource allocation in the hierarchical medical treatment system in China	Patient diagnostic characteristics and hospital tiers
McRae et al (2020) [24]	To present a clinical decision-support system and mobile app to assist in COVID severity assessment, management, and care	Resource allocation during COVID in New York, with validation data sets from Wuhan, China	Outpatient score (age, gender, diabetes, cardiovascular comorbidities, and systolic blood pressure) and biomarker score (C-reactive protein, procalcitonin, and age)
Bednarski et al (2021) [21]	To study how reinforcement learning and deep-learning models can facilitate the redistribution of medical equipment during pandemics	Pandemics in the context of COVID	COVID risk factors by region, COVID mortality by region, and current demand for medical equipment

Summary of the Characteristics of Studies Involving Reviews and Theoretical Discussions

The characteristics of studies involving reviews and theoretical discussions are summarized in Table 3. Of the 22 included studies, 13 were theoretical discussions, qualitative studies, or review studies [31-43]. Of those studies, 8 studies were qualitative discussions on the COVID-19 pandemic

[31,33,34,36,38,39,41,43], with 2 in a Chinese context [34,43] and the rest in a global situation. The remaining 5 studies focused on other situations, with 1 focusing on resource allocation in intensive care units and hospital stay [40], 1 on disease outbreaks and disasters [33], 1 on diabetic retinopathy screening [42], 1 on human resource allocation in health systems [35], and 1 on medical information digitalization [37].

Table 3. Characteristics of the included studies involving theoretical discussions, qualitative studies, or review studies.

Reference	Objective	Resource allocation situation	Reviewed/discussed methods for the application of AI ^a during the COVID-19 pandemic
Rajkomar et al (2018) [40]	To explore how model design, biases in data, and interactions of model predictions with clinicians and patients exacerbate health inequalities	Intensive care unit and in-hospital stay length	<ul style="list-style-type: none"> Suggested that future AI models for health care resource distribution should include principles of distributive justice.
Laudanski et al (2020) [36]	To analyze the applications of AI during COVID using the WHO ^b framework of pandemic evolution	Global COVID-19 pandemic	<ul style="list-style-type: none"> Reviewed cases in Italy where AI was used in studying computed tomography scans for COVID prognosis, and suggested that AI-driven scans can help predict prognosis and therefore allow better resource distribution. Discussed AI-driven triage based on patient characteristics and AI-supported health resource allocation and ethics.
Adly et al (2020) [31]	To discuss the potential of using AI to prevent and control COVID	Global COVID-19 pandemic	<ul style="list-style-type: none"> Suggested that the application of AI was valuable in medical resource distribution that included the parameters of patients and the pandemic.
Bernardo et al (2020) [33]	To present approaches for using technology to facilitate resource distribution in disasters and outbreaks	Disasters and disease outbreaks	<ul style="list-style-type: none"> Found that data collected from crowdsourcing and the human-technology interface could be used as data sources.
Neves et al (2020) [38]	To discuss the basic principles of medical resource allocation choices during COVID	Global COVID-19 pandemic	<ul style="list-style-type: none"> Discussed rationalization of care, medical and team conflict, modeling of the pandemic, and application of AI. Explored the use of AI as a support tool to streamline inventory control and standardize resource distribution.
Xie et al (2020) [42]	To present an overview of the application of AI technology in ophthalmology, with a focus on deep-learning systems	Diabetic retinopathy screening	<ul style="list-style-type: none"> Reviewed empirical considerations behind the formation of successful screening programs. Examined potential methods for health economics and safety analyses that can assess concerns regarding AI-based screening.
Zou et al (2020) [43]	To present the COVID response of Shenzhen, China and discuss the potential of a successful model for COVID prevention and control	COVID-19 pandemic in Shenzhen, China	<ul style="list-style-type: none"> Reviewed methods applied by Shenzhen, including early action and centralized response, care for vulnerable persons, community response teams, and technology. Discussed the integration of information technology in Shenzhen's response, including mobile technology, big data, and AI.
Basit et al (2021) [32]	To discuss the data sharing and collection process and the ethical considerations around pandemic data	Global COVID-19 pandemic	<ul style="list-style-type: none"> Discussed the required data, failures and challenges in obtaining pandemic data, success in data access, model creation using data, and ethical challenges associated with data access during the COVID-19 pandemic. Discussed the application of AI in the allocation of intensive care resources and ventilators.
Huang et al (2021) [34]	To investigate China's health informatization, especially during the COVID-19 pandemic	COVID-19 pandemic in China	<ul style="list-style-type: none"> Discussed the development of China's health informatization from 5 perspectives: health information infrastructure, information technology applications, financial and intellectual investment, health resource allocation, and the standard system.
Jain et al (2021) [35]	To discuss the implications of AI for employability by analyzing issues in the health care sector	Human resources in health systems	<ul style="list-style-type: none"> Displayed hierarchical relationships between employability and a range of characteristics. Discussed measures that could potentially enhance employability in the health care sector through AI.
Lu et al (2021) [37]	To establish barriers that affect medical information digitalization innovation and development through interviews and a literature review	Medical information digitalization	<ul style="list-style-type: none"> Applied the importance-resistance analysis model and identified the resistant factors, including data sharing, infrastructure, regulation, and operations in the context of data privacy. Proposed several ways to overcome these limitations, including transparency regulation and infrastructure building.

Reference	Objective	Resource allocation situation	Reviewed/discussed methods for the application of AI ^a during the COVID-19 pandemic
Pereira et al (2021) [39]	To present interindividual variability and the roles it plays in the variability of COVID presentation and susceptibility.	Global COVID-19 pandemic	<ul style="list-style-type: none"> Reviewed the biological differences that contribute to variability in COVID manifestation. Reviewed efforts to use AI to integrate digital data to enable the identification of high-risk COVID-19 patients.
Röösli et al (2021) [41]	To discuss possible bias in the application of AI during the COVID-19 pandemic	Global COVID-19 pandemic	<ul style="list-style-type: none"> Discussed how COVID exacerbated racial and socioeconomic disparities. Explored how an AI-informed resource allocation strategy can be influenced by biases.

^aAI: artificial intelligence.

^bWHO: World Health Organization.

Summary of the Policy Implications of the Selected Studies

The policy implications of studies on model development are relevant on 2 levels: (1) health system level [21,25,26,29] and (2) hospital level [22-24,27,28], corresponding to situations where the models were applied. Detailed policy implications of the included studies on model development are summarized

in Table 4. The qualitative and review studies focused largely on 2 issues: (1) how AI can promote the efficacy of resource allocation [21,32,34-37,39,42,43] and (2) the ethics and equality issues associated with using AI systems [38,40,41]. One study highlighted the lack of AI studies on resource distribution during COVID-19 [31]. Table 5 summarizes the policy implications of these studies.

Table 4. Policy relevance of the included studies on model development and validation.

Reference	Policy relevance
Rosas et al (2013) [25]	<ul style="list-style-type: none"> Divided municipalities in Brazil into quartiles of health care financial needs. Proposed that the selection of input variables should consider the vulnerability of the population, the true representation of the factors of need, political choice, and the availability of reliable data.
Belciug & Gorunescu (2015) [22]	<ul style="list-style-type: none"> Provided tools to estimate the appropriate parameters for optimal resource utilization. Enabled the hospital manager to simulate scenarios to make the near-best decision.
Gartner & Padman (2015) [23]	<ul style="list-style-type: none"> Provided decision-makers with information on admission and scheduling decisions. Offered an approach to integrate and analyze the financial objectives of health care delivery.
Velez et al (2016) [26]	<ul style="list-style-type: none"> Facilitated the management of multidisciplinary information with the entire range of determinants of a specific context. Provided enough flexibility to allow the exploration of different complex circumstances in health planning.
Xu et al (2018) [27]	<ul style="list-style-type: none"> Reduced costs by making doctors mobile. Addressed personal preferences, such as treatment time and the professional level of doctors.
Yousefi et al (2018) [28]	<ul style="list-style-type: none"> Decreased the average length of stay in this emergency department case study by 14%. Provided a framework to efficiently combine simulation and metamodels in the health care industry.
Zhang et al (2018) [29]	<ul style="list-style-type: none"> Facilitated decision-making to divide patients under different conditions into different levels of hospitals in the hierarchical medical treatment system.
McRae et al (2020) [24]	<ul style="list-style-type: none"> Supported the validity of a clinical decision support system and mobile app Provided tools to be deployed to community clinics and sites for decision support.
Bednarski et al (2021) [21]	<ul style="list-style-type: none"> Facilitated officials managing future public health crises. Improved algorithm performance for future applications.

Table 5. Policy relevance of the included studies involving theoretical discussions, qualitative studies, or review studies.

Reference	Policy relevance
Rajkomar et al (2018) [40]	<ul style="list-style-type: none"> Proposed that the principles of distributive justice be incorporated into model design, deployment, and evaluation.
Laudanski et al (2020) [36]	<ul style="list-style-type: none"> Suggested that AI^a can couple outbreak data with measures of potential demand and direct supplies more efficiently.
Adly et al (2020) [31]	<ul style="list-style-type: none"> Found that no study had been published on the application of AI in medical resource distribution during the COVID-19 pandemic as of 2020 and that such studies are required to inform policy decisions.
Bernardo et al (2020) [33]	<ul style="list-style-type: none"> Suggested that automation by AI and machine learning can further our abilities in predictive analytics.
Neves et al (2020) [38]	<ul style="list-style-type: none"> Emphasized that the ethical values for the rationing of health resources in an epidemic should converge with basic ethical values and that transparency is essential to ensure public trust.
Xie et al (2020) [42]	<ul style="list-style-type: none"> Proposed that technical feasibility and patient acceptability must be assessed for AI to be deployed in real-world settings, and that health professionals' acceptance and interpretability of AI-based screening strategies must also be assessed.
Zou et al (2020) [43]	<ul style="list-style-type: none"> Proposed that the model adopted in Shenzhen, including multisectoral coordination, proactive contact tracing and testing, timely isolation and treatment, hospital infection control, effective community management, and prompt information dissemination, could be a potential model for other cities around the world for containing the pandemic.
Basit et al (2021) [32]	<ul style="list-style-type: none"> Proposed that informaticians globally should continue collecting, recording, and analyzing data with the intent of gathering new knowledge and translating it into a better, faster, and more successful response to the next pandemic. Suggested that professionals must come together to develop ways to collect, standardize, and disseminate the data needed to make necessary decisions.
Huang et al (2021) [34]	<ul style="list-style-type: none"> Suggested that China's health informatization needs to strengthen top-level design, increase investment and training, upgrade health infrastructure and information technology applications, and improve internet-based health care services.
Jain et al (2021) [35]	<ul style="list-style-type: none"> Proposed that an AI intervention could impact the employability of the workforce through operational and training changes, and therefore impact human resource distribution in health.
Lu et al (2021) [37]	<ul style="list-style-type: none"> Provided a basis for the future development directions of medical information digitalization and its impacts on health care and health systems.
Pereira et al (2021) [39]	<ul style="list-style-type: none"> Suggested that predicting which COVID-19 patients will develop progressive diseases that require hospitalization has important implications for clinical trials targeting outpatients.
Röösli et al (2021) [41]	<ul style="list-style-type: none"> Proposed that transparency in reporting of AI algorithms is necessary to understand intended predictions, target populations, hidden biases, and class imbalance problems.

^aAI: artificial intelligence.

Case Study Comparison: China and Brazil

China and Brazil are both developing countries with a similar per capita gross domestic product (China: US \$10,435 and Brazil: US \$6797) [44]. During the COVID-19 pandemic, Brazil has had one of the highest national overall cases and mortalities, as well as per capita cases and mortalities, with 29.5 million cases and 656,000 deaths as of March 2022 [45]. China has had one of the lowest per capita infection rates in the world, with a total of 124,000 cases and 4636 deaths as of March 2022 [45]. Given the similarity between the 2 countries in economic development and the enormous difference in COVID cases and mortalities, the resource distribution situation in the 2 countries is worth exploring.

Rosas et al [25] proposed a financial resource allocation algorithm for the public hospital system in Brazil based on mortality, socioeconomic characteristics, and income inequality. They argued that the choice of input variables for health care

policymaking should consider the vulnerability of the population to being manipulated by those who manage public policy, the true representation of the factors of need, exemption from the process of political choice, and the availability of reliable data. The focus of the model was regional economic characteristics.

Zhang et al [29] proposed a model for the allocation of medical resources and tier classification of patients in China's health system, with the input variables of patient characteristics and hospital tiers, and a focus on differentiation into different tiers based on patients' disease severity. Xu et al [27] proposed a health resource allocation model for the allocation of doctors and other medical resources in a public hospital system in China that considered the distribution of medical stations, the professional level of doctors (salary and seniority), patient preferences and illness severity, medical cost, and revenue.

Overall, the allocation of medical resources based on the models from the 3 studies demonstrated that the key considerations

proposed by studies from China were the hospital tier system, the professional level of doctors, the geographical distribution of medical resources, and cost-effectiveness [27,29]. However, the model proposed for Brazil focused on the regional economic situation [25].

Discussion

Principal Findings

In this review, we compiled evidence on the application of AI in health resource distribution, especially regarding COVID-related policy. After synthesizing 22 articles, we found that AI-based models were proposed at both hospital (secondary care in inpatient settings) and health system (public health) levels and that theoretical discussions and reviews focused on the potential for AI to improve the efficacy of resource distribution and on the ethics of applying AI in health resource distribution. Two major themes emerged from the review. First, we found that AI-informed resource distribution strategies are impactful for health access and equality. Second, the approaches can be categorized ideologically into revisionist and conservative groups.

Impact of an AI-Informed Resource Distribution Strategy on Health Access and Equality

AI and machine learning have considerable potential to improve efficacy in resource distribution, especially during emergencies, such as the COVID-19 pandemic, where quick decisions are required based on evolving situations [34,39,43]. For example, health informatization, particularly digital contact tracing and AI-informed response design, played an instrumental role in responding to COVID in China and helped local governments to improve efficacy in allocating limited resources [34,43]. AI can also be used to interpret diagnostic results and patient characteristics in order to predict disease progression and allocation of medicines, hospital beds, and medical professionals at the hospital level [21,39].

However, very large amounts of data are necessary for AI algorithms to make reliable and evidence-based decisions [46]. Health care institutions globally must therefore collect, record, and analyze data. This will help policymakers gather novel insights and translate the data into a prompt, equal, coordinated, and more successful response to the next pandemic [32,47]. As such, data collection must be institutionalized. The disparity in data collection capacity potentially exacerbates the gap in decision-making quality between countries [48,49]. For example, from the literature, China's information infrastructure and data-sharing agreements expedited the data-gathering process, a possible consequence of the centralized government system that facilitated gathering data, which in turn made the data set larger and more comprehensive [48]. In contrast, a selected study showed that Brazil's decentralized government system, with heterogeneous policies on data privacy and data sharing, made the collection and consolidation of data difficult [49]. However, caution should be taken in interpreting those results, as there is no evidence that the studies selected here are representative of the real situation in China or Brazil.

The included articles highlighted the importance of distributive justice and transparency in AI model design. The analysis conducted by Rajkomar et al emphasized that machine learning systems should be used proactively to advance health equality [40]. They proposed that distributive justice should be a core principle in AI models, including during the design, deployment, and evaluation processes. This perspective would include equality in patient outcomes, performance for every sociodemographic group, and resource allocation for each group. As Neves et al noted, resource allocation by AI and in emergencies should build on basic ethical values, including the equal value of people, instrumental value, and priority for critical situations. Transparency is the key to gaining trust when distributing resources [38].

Revisionist and Conservative Approaches in AI-Derived Resource Distribution

The build-up of AI models and implementation plans can be broadly categorized into revisionist and conservative approaches. In revisionist approaches, the models aim to revise the disparity in resource distribution by actively correcting the biases in previous decision-making processes. For example, the models proposed by Rosas et al [25] for financial resource allocation in Brazil emphasized consideration of income inequality, vulnerable populations, political choices, and the availability of reliable data. In conservative approaches, the models rely on traditional metrics, including supply and demand, profitability, and, perhaps most notably, previous decisions. This was demonstrated in a proposed model for the allocation of medical resources and tier classification of patients in China's health system by Zhang et al [29], where the input variables were patients' characteristics and hospital tiers, and a model suggested by Xu et al [27] for the allocation of doctors and other medical resources in a public hospital system in China, where the input variables included the distribution of medical stations, the professional level of doctors, patient preferences and illness severity, medical cost, and revenue. Doctor expertise, patient characteristics, hospital tier, and location are common variables in human decision-making, but AI has the potential to analyze the data more thoroughly.

However, despite the revisionist model proposed by Brazilian academics [25], health inequality is a prevailing issue in Brazil across states and social classes, both before [50] and during the COVID-19 pandemic [51]. Health inequality in Brazil increased across states from 1990 to 2016 [43]. Comparatively, the health care access and quality index in China was higher than that in Brazil in 2016, suggesting better equality and health care access in China [52]. However, due to the limitation of the research method, this study could not show the policymaking processes in both countries. From the selected studies alone, we observed that although proposing revisionist AI models to address health inequality should be encouraged, the application and practicality of using those models to inform health policy decisions and improve inequality should also be important considerations for researchers.

Strengths and Limitations

This is one of the first reviews to incorporate all available evidence qualitatively and provide a comprehensive picture of

the model development and theoretical discussion on AI in medical resource distribution. Our results contribute to the ongoing discussion of applying AI in medical resource distribution and add novel insights into the social and ethical implications. Nonetheless, this study has several limitations. First, due to the scope of the study, we focused on published journal articles but did not examine policy documents or grey literature. This could have led to incompleteness in the collected information. Further studies could examine policy statements and grey literature to better understand intercountry differences. Second, we included only articles published in English and therefore might have overlooked publications in other languages. Third, there are potential sources of meaningful heterogeneity in this scoping review, including the diverse use of AI technologies, different study designs, and different locations. The analyses in this study could be affected by such heterogeneities. Fourth, this study is a qualitative overview of the general application of AI in health care resource distribution and is exploratory. We did not compare different levels of resource distribution and distinguish various machine learning methods in detail. Further studies are needed to explore and

contrast different AI approaches at various resource distribution levels in detail. Lastly, due to the availability of evidence, we only compared studies from China and Brazil. We were only able to compare the differences between the 2 countries based on a few studies, which could not represent the real situation in either country. The comparison should be interpreted as exploratory and demonstrative.

Conclusions

This scoping review synthesized evidence on the application of AI in health resource distribution, particularly during the COVID pandemic. The included studies suggested that AI and machine learning have high potentials to improve efficacy in resource distribution, especially during sudden and evolving situations. A coordinated and continuous data sharing and collecting mechanism is needed for better data input so that AI can make reliable and evidence-based decisions. Various issues, including health inequality, distributive justice, and transparency, should be considered when deploying AI models. Such considerations are required for implementing revisionist AI models that can correct distribution inequality in actual policy processes.

Conflicts of Interest

None declared.

References

1. van der Schaar M, Alaa AM, Floto A, Gimson A, Scholtes S, Wood A, et al. How artificial intelligence and machine learning can help healthcare systems respond to COVID-19. *Mach Learn* 2021 Dec 09;110(1):1-14 [FREE Full text] [doi: [10.1007/s10994-020-05928-x](https://doi.org/10.1007/s10994-020-05928-x)] [Medline: [33318723](https://pubmed.ncbi.nlm.nih.gov/33318723/)]
2. Mintz Y, Brodie R. Introduction to artificial intelligence in medicine. *Minim Invasive Ther Allied Technol* 2019 Apr 27;28(2):73-81. [doi: [10.1080/13645706.2019.1575882](https://doi.org/10.1080/13645706.2019.1575882)] [Medline: [30810430](https://pubmed.ncbi.nlm.nih.gov/30810430/)]
3. Amisha, Malik P, Pathania M, Rathaur V. Overview of artificial intelligence in medicine. *J Family Med Prim Care* 2019 Jul;8(7):2328-2331 [FREE Full text] [doi: [10.4103/jfmpc.jfmpc_440_19](https://doi.org/10.4103/jfmpc.jfmpc_440_19)] [Medline: [31463251](https://pubmed.ncbi.nlm.nih.gov/31463251/)]
4. Le HH, Edman H, Honda Y, Kushima M, Yamazaki T, Araki K, et al. Fast Generation of Clinical Pathways including Time Intervals in Sequential Pattern Mining on Electronic Medical Record Systems. 2017 Presented at: International Conference on Computational Science and Computational Intelligence (CSCI); December 14-16, 2017; Las Vegas, NV, USA. [doi: [10.1109/CSCI.2017.300](https://doi.org/10.1109/CSCI.2017.300)]
5. Yu M, He S, Wu D, Zhu H, Webster C. Examining the Multi-Scalar Unevenness of High-Quality Healthcare Resources Distribution in China. *Int J Environ Res Public Health* 2019 Aug 07;16(16):2813 [FREE Full text] [doi: [10.3390/ijerph16162813](https://doi.org/10.3390/ijerph16162813)] [Medline: [31394765](https://pubmed.ncbi.nlm.nih.gov/31394765/)]
6. Hulshof PJH, Kortbeek N, Boucherie RJ, Hans EW, Bakker PJM. Taxonomic classification of planning decisions in health care: a structured review of the state of the art in OR/MS. *Health Systems* 2017 Dec 19;1(2):129-175. [doi: [10.1057/hs.2012.18](https://doi.org/10.1057/hs.2012.18)]
7. Luengo-Oroz M, Hoffmann Pham K, Bullock J, Kirkpatrick R, Luccioni A, Rubel S, et al. Artificial intelligence cooperation to support the global response to COVID-19. *Nat Mach Intell* 2020 May 22;2(6):295-297. [doi: [10.1038/s42256-020-0184-3](https://doi.org/10.1038/s42256-020-0184-3)]
8. Zhou Y, Wang F, Tang J, Nussinov R, Cheng F. Artificial intelligence in COVID-19 drug repurposing. *The Lancet Digital Health* 2020 Dec;2(12):e667-e676. [doi: [10.1016/s2589-7500\(20\)30192-8](https://doi.org/10.1016/s2589-7500(20)30192-8)]
9. Cheng F, Desai RJ, Handy DE, Wang R, Schneeweiss S, Barabási AL, et al. Network-based approach to prediction and population-based validation of in silico drug repurposing. *Nat Commun* 2018 Jul 12;9(1):2691 [FREE Full text] [doi: [10.1038/s41467-018-05116-5](https://doi.org/10.1038/s41467-018-05116-5)] [Medline: [30002366](https://pubmed.ncbi.nlm.nih.gov/30002366/)]
10. Jin C, Chen W, Cao Y, Xu Z, Tan Z, Zhang X, et al. Development and evaluation of an artificial intelligence system for COVID-19 diagnosis. *Nat Commun* 2020 Oct 09;11(1):5088 [FREE Full text] [doi: [10.1038/s41467-020-18685-1](https://doi.org/10.1038/s41467-020-18685-1)] [Medline: [33037212](https://pubmed.ncbi.nlm.nih.gov/33037212/)]
11. Mei X, Lee H, Diao K, Huang M, Lin B, Liu C, et al. Artificial intelligence-enabled rapid diagnosis of patients with COVID-19. *Nat Med* 2020 Aug 19;26(8):1224-1228 [FREE Full text] [doi: [10.1038/s41591-020-0931-3](https://doi.org/10.1038/s41591-020-0931-3)] [Medline: [32427924](https://pubmed.ncbi.nlm.nih.gov/32427924/)]

12. Hu Z, Ge Q, Li S, Jin L, Xiong M. Artificial Intelligence Forecasting of Covid-19 in China. arXiv. 2020. URL: <https://arxiv.org/abs/2002.07112> [accessed 2023-01-15]
13. Reeve J, Richardson C. Artificial intelligence (AI) and Britons health: how can AI help to health in resource-based situations? *International Journal of Modern Engineering Technologies* 2020;2(2).
14. Kong X, Ai B, Kong Y, Su L, Ning Y, Howard N, et al. Artificial intelligence: a key to relieve China's insufficient and unequally-distributed medical resources. *Am J Transl Res* 2019;11(5):2632-2640 [FREE Full text] [Medline: [31217843](#)]
15. Tan-Torres Edejer T, Hanssen O, Mirelman A, Verboom P, Lolong G, Watson OJ, et al. Projected health-care resource needs for an effective response to COVID-19 in 73 low-income and middle-income countries: a modelling study. *Lancet Glob Health* 2020 Nov;8(11):e1372-e1379 [FREE Full text] [doi: [10.1016/S2214-109X\(20\)30383-1](#)] [Medline: [32918872](#)]
16. Ji Y, Ma Z, Peppelenbosch MP, Pan Q. Potential association between COVID-19 mortality and health-care resource availability. *Lancet Glob Health* 2020 Apr;8(4):e480 [FREE Full text] [doi: [10.1016/S2214-109X\(20\)30068-1](#)] [Medline: [32109372](#)]
17. Lane H, Sarkies M, Martin J, Haines T. Equity in healthcare resource allocation decision making: A systematic review. *Soc Sci Med* 2017 Feb;175:11-27. [doi: [10.1016/j.socscimed.2016.12.012](#)] [Medline: [28043036](#)]
18. Kang J, Michels A, Lyu F, Wang S, Agbodo N, Freeman VL, et al. Rapidly measuring spatial accessibility of COVID-19 healthcare resources: a case study of Illinois, USA. *Int J Health Geogr* 2020 Sep 14;19(1):36 [FREE Full text] [doi: [10.1186/s12942-020-00229-x](#)] [Medline: [32928236](#)]
19. Arksey H, O'Malley L. Scoping studies: towards a methodological framework. *International Journal of Social Research Methodology* 2005 Feb;8(1):19-32. [doi: [10.1080/1364557032000119616](#)]
20. Tricco AC, Lillie E, Zarin W, O'Brien KK, Colquhoun H, Levac D, et al. PRISMA Extension for Scoping Reviews (PRISMA-ScR): Checklist and Explanation. *Ann Intern Med* 2018 Oct 02;169(7):467-473 [FREE Full text] [doi: [10.7326/M18-0850](#)] [Medline: [30178033](#)]
21. Bednarski B, Singh A, Jones W. On collaborative reinforcement learning to optimize the redistribution of critical medical supplies throughout the COVID-19 pandemic. *J Am Med Inform Assoc* 2021 Mar 18;28(4):874-878 [FREE Full text] [doi: [10.1093/jamia/ocaa324](#)] [Medline: [33295626](#)]
22. Belciug S, Gorunescu F. Improving hospital bed occupancy and resource utilization through queuing modeling and evolutionary computation. *J Biomed Inform* 2015 Feb;53:261-269 [FREE Full text] [doi: [10.1016/j.jbi.2014.11.010](#)] [Medline: [25433363](#)]
23. Gartner D, Padman R. Improving Hospital-Wide Early Resource Allocation through Machine Learning. *Stud Health Technol Inform* 2015;216:315-319. [Medline: [26262062](#)]
24. McRae MP, Dapkins IP, Sharif I, Anderman J, Fenyo D, Sinokrot O, et al. Managing COVID-19 With a Clinical Decision Support Tool in a Community Health Network: Algorithm Development and Validation. *J Med Internet Res* 2020 Aug 24;22(8):e22033 [FREE Full text] [doi: [10.2196/22033](#)] [Medline: [32750010](#)]
25. Rosas MA, Bezerra AFB, Duarte-Neto PJ. Use of artificial neural networks in applying methodology for allocating health resources. *Rev Saude Publica* 2013 Feb;47(1):128-36; discussion 136 [FREE Full text] [doi: [10.1590/s0034-89102013000100017](#)] [Medline: [23703139](#)]
26. Velez AFJ, Rosas JDC, Fierro JMM. Geospatial model e-health planning collective intelligence. 2016 Presented at: Third International Conference on eDemocracy & eGovernment (ICEDEG); March 30, 2016-April 01, 2016; Sangolqui, Ecuador. [doi: [10.1109/ICEDEG.2016.7461708](#)]
27. Xu Y, Liu S, Wang B. Research on Computational Intelligence in Medical Resource Allocation Based on Mass Customization. *Journal of Universal Computer Science* 2018;24(6):753-774.
28. Yousefi M, Yousefi M, Ferreira RPM, Kim JH, Fogliatto FS. Chaotic genetic algorithm and Adaboost ensemble metamodeling approach for optimum resource planning in emergency departments. *Artif Intell Med* 2018 Jan;84:23-33. [doi: [10.1016/j.artmed.2017.10.002](#)] [Medline: [29054572](#)]
29. Zhang R, Xing Y, Wang J, Shang X, Zhu X. A Novel Multiattribute Decision-Making Method Based on Point Choquet Aggregation Operators and Its Application in Supporting the Hierarchical Medical Treatment System in China. *Int J Environ Res Public Health* 2018 Aug 10;15(8):1718 [FREE Full text] [doi: [10.3390/ijerph15081718](#)] [Medline: [30103454](#)]
30. Gartner D, Padman R. Improving Hospital-Wide Early Resource Allocation through Machine Learning. In: *MEDINFO 2015: eHealth-enabled Health*. Amsterdam, the Netherlands: IOS Press; 2015:315-319.
31. Adly AS, Adly AS, Adly MS. Approaches Based on Artificial Intelligence and the Internet of Intelligent Things to Prevent the Spread of COVID-19: Scoping Review. *J Med Internet Res* 2020 Aug 10;22(8):e19104 [FREE Full text] [doi: [10.2196/19104](#)] [Medline: [32584780](#)]
32. Basit MA, Lehmann CU, Medford RJ. Managing Pandemics with Health Informatics: Successes and Challenges. *Yearb Med Inform* 2021 Aug 21;30(1):17-25 [FREE Full text] [doi: [10.1055/s-0041-1726478](#)] [Medline: [33882594](#)]
33. Bernardo TM, Perez Gutierrez E, Hachborn GF, Forrest RO, Sobkowich KE. Innovating at the human-technology interface in disasters and disease outbreaks. *Rev Sci Tech* 2020 Aug 01;39(2):491-501 [FREE Full text] [doi: [10.20506/rst.39.2.3100](#)] [Medline: [33046926](#)]

34. Huang M, Wang J, Nicholas S, Maitland E, Guo Z. Development, Status Quo, and Challenges to China's Health Informatization During COVID-19: Evaluation and Recommendations. *J Med Internet Res* 2021 Jun 17;23(6):e27345 [FREE Full text] [doi: [10.2196/27345](https://doi.org/10.2196/27345)] [Medline: [34061761](https://pubmed.ncbi.nlm.nih.gov/34061761/)]
35. Jain M, Goel A, Sinha S, Dhir S. Employability implications of artificial intelligence in healthcare ecosystem: responding with readiness. *FS* 2021 Jan 04;23(1):73-94. [doi: [10.1108/fs-04-2020-0038](https://doi.org/10.1108/fs-04-2020-0038)]
36. Laudanski K, Shea G, DiMeglio M, Rastrepo M, Solomon C. What Can COVID-19 Teach Us about Using AI in Pandemics? *Healthcare (Basel)* 2020 Dec 01;8(4):527 [FREE Full text] [doi: [10.3390/healthcare8040527](https://doi.org/10.3390/healthcare8040527)] [Medline: [33271960](https://pubmed.ncbi.nlm.nih.gov/33271960/)]
37. Lu W, Tsai I, Wang K, Tang T, Li K, Ke Y, et al. Innovation Resistance and Resource Allocation Strategy of Medical Information Digitalization. *Sustainability* 2021 Jul 14;13(14):7888. [doi: [10.3390/su13147888](https://doi.org/10.3390/su13147888)]
38. Neves NMBC, Bitencourt FBCSN, Bitencourt AGV. Ethical dilemmas in COVID-19 times: how to decide who lives and who dies? *Rev. Assoc. Med. Bras* 2020;66(suppl 2):106-111. [doi: [10.1590/1806-9282.66.s2.106](https://doi.org/10.1590/1806-9282.66.s2.106)]
39. Pereira NL, Ahmad F, Byku M, Cummins NW, Morris AA, Owens A, et al. COVID-19: Understanding Inter-Individual Variability and Implications for Precision Medicine. *Mayo Clin Proc* 2021 Feb;96(2):446-463 [FREE Full text] [doi: [10.1016/j.mayocp.2020.11.024](https://doi.org/10.1016/j.mayocp.2020.11.024)] [Medline: [33549263](https://pubmed.ncbi.nlm.nih.gov/33549263/)]
40. Rajkomar A, Hardt M, Howell MD, Corrado G, Chin MH. Ensuring Fairness in Machine Learning to Advance Health Equity. *Ann Intern Med* 2018 Dec 18;169(12):866-872 [FREE Full text] [doi: [10.7326/M18-1990](https://doi.org/10.7326/M18-1990)] [Medline: [30508424](https://pubmed.ncbi.nlm.nih.gov/30508424/)]
41. Rössli E, Rice B, Hernandez-Boussard T. Bias at warp speed: how AI may contribute to the disparities gap in the time of COVID-19. *J Am Med Inform Assoc* 2021 Jan 15;28(1):190-192 [FREE Full text] [doi: [10.1093/jamia/ocaa210](https://doi.org/10.1093/jamia/ocaa210)] [Medline: [32805004](https://pubmed.ncbi.nlm.nih.gov/32805004/)]
42. Xie Y, Gunasekeran DV, Balaskas K, Keane PA, Sim DA, Bachmann LM, et al. Health Economic and Safety Considerations for Artificial Intelligence Applications in Diabetic Retinopathy Screening. *Transl Vis Sci Technol* 2020 Apr;9(2):22 [FREE Full text] [doi: [10.1167/tvst.9.2.22](https://doi.org/10.1167/tvst.9.2.22)] [Medline: [32818083](https://pubmed.ncbi.nlm.nih.gov/32818083/)]
43. Zou H, Shu Y, Feng T. How Shenzhen, China avoided widespread community transmission: a potential model for successful prevention and control of COVID-19. *Infect Dis Poverty* 2020 Jul 10;9(1):89 [FREE Full text] [doi: [10.1186/s40249-020-00714-2](https://doi.org/10.1186/s40249-020-00714-2)] [Medline: [32650840](https://pubmed.ncbi.nlm.nih.gov/32650840/)]
44. GDP per capita (current US\$). World Bank. URL: <https://data.worldbank.org/indicator/NY.GDP.PCAP.CD> [accessed 2022-03-16]
45. Novel Coronavirus (COVID-19) Cases, provided by JHU CSSE. GitHub. URL: <https://github.com/CSSEGISandData/COVID-19> [accessed 2022-03-16]
46. Chen M, Decary M. Artificial intelligence in healthcare: An essential guide for health leaders. *Healthc Manage Forum* 2020 Jan;33(1):10-18. [doi: [10.1177/0840470419873123](https://doi.org/10.1177/0840470419873123)] [Medline: [31550922](https://pubmed.ncbi.nlm.nih.gov/31550922/)]
47. Wirtz BW, Weyerer JC, Geyer C. Artificial Intelligence and the Public Sector—Applications and Challenges. *International Journal of Public Administration* 2018 Jul 24;42(7):596-615. [doi: [10.1080/01900692.2018.1498103](https://doi.org/10.1080/01900692.2018.1498103)]
48. Zhang L, Wang H, Li Q, Zhao M, Zhan Q. Big data and medical research in China. *BMJ* 2018 Feb 05;360:j5910 [FREE Full text] [doi: [10.1136/bmj.j5910](https://doi.org/10.1136/bmj.j5910)] [Medline: [29437562](https://pubmed.ncbi.nlm.nih.gov/29437562/)]
49. Ferreira L, Okano MT, Aguiar F, dos Santos HDCL, Ursini E. A panorama of the implementation of the General Law for the Protection of Personal Data (LGPD) in Brazil: an exploratory survey. 2022 Presented at: 12th Annual Computing and Communication Workshop and Conference (CCWC); January 26-29, 2022; Las Vegas, NV, USA. [doi: [10.1109/CCWC54503.2022.9720879](https://doi.org/10.1109/CCWC54503.2022.9720879)]
50. Santos JAF. Classe Social, território e desigualdade de saúde no Brasil. *Saude Soc* 2018 Jun 25;27(2):556-572. [doi: [10.1590/s0104-12902018170889](https://doi.org/10.1590/s0104-12902018170889)]
51. Malta M, Murray L, da Silva CMFP, Strathdee SA. Coronavirus in Brazil: The heavy weight of inequality and unsound leadership. *EClinicalMedicine* 2020 Aug;25:100472 [FREE Full text] [doi: [10.1016/j.eclinm.2020.100472](https://doi.org/10.1016/j.eclinm.2020.100472)] [Medline: [32754684](https://pubmed.ncbi.nlm.nih.gov/32754684/)]
52. GBD 2016 Healthcare Access and Quality Collaborators. Measuring performance on the Healthcare Access and Quality Index for 195 countries and territories and selected subnational locations: a systematic analysis from the Global Burden of Disease Study 2016. *Lancet* 2018 Jun 02;391(10136):2236-2271 [FREE Full text] [doi: [10.1016/S0140-6736\(18\)30994-2](https://doi.org/10.1016/S0140-6736(18)30994-2)] [Medline: [29893224](https://pubmed.ncbi.nlm.nih.gov/29893224/)]

Abbreviations

AI: artificial intelligence

Edited by B Malin, K El Emam; submitted 31.03.22; peer-reviewed by W Perveen, H Mehdizadeh, D Vurmaz, ER Khalilian, D Gartner; comments to author 14.06.22; revised version received 29.12.22; accepted 06.01.23; published 30.01.23.

Please cite as:

Wu H, Lu X, Wang H

The Application of Artificial Intelligence in Health Care Resource Allocation Before and During the COVID-19 Pandemic: Scoping Review

JMIR AI 2023;2:e38397

URL: <https://ai.jmir.org/2023/1/e38397>

doi: [10.2196/38397](https://doi.org/10.2196/38397)

PMID: [27917920](https://pubmed.ncbi.nlm.nih.gov/27917920/)

©Hao Wu, Xiaoyu Lu, Hanyu Wang. Originally published in JMIR AI (<https://ai.jmir.org>), 30.01.2023. This is an open-access article distributed under the terms of the Creative Commons Attribution License (<https://creativecommons.org/licenses/by/4.0/>), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR AI, is properly cited. The complete bibliographic information, a link to the original publication on <https://www.ai.jmir.org/>, as well as this copyright and license information must be included.

Original Paper

Predicting Treatment Interruption Among People Living With HIV in Nigeria: Machine Learning Approach

Matthew-David Ogbechie^{1*}, BSc; Christa Fischer Walker^{2*}, PhD; Mu-Tien Lee³, MSc; Amina Abba Gana¹, MSc; Abimbola Oduola¹, PhD; Augustine Idemudia¹, MPH, MD; Matthew Edor¹, BSc; Emily Lark Harris⁴, MA; Jessica Stephens⁵, MPH; Xiaoming Gao³, MSc; Pai-Lien Chen³, PhD; Navindra Etwaroo Persaud^{2*}, MBBS, MPH, PhD

¹FHI 360, Abuja, Nigeria

²FHI 360, Washington, DC, United States

³FHI 360, Durham, NC, United States

⁴United States Agency for International Development, Dar es Salaam, United Republic of Tanzania

⁵United States Agency for International Development, Washington, DC, United States

*these authors contributed equally

Corresponding Author:

Navindra Etwaroo Persaud, MBBS, MPH, PhD

FHI 360

1825 Connecticut Ave NW

Washington, DC, 20009

United States

Phone: 1 2028848017

Email: npersaud@fhi360.org

Abstract

Background: Antiretroviral therapy (ART) has transformed HIV from a fatal illness to a chronic disease. Given the high rate of treatment interruptions, HIV programs use a range of approaches to support individuals in adhering to ART and in re-engaging those who interrupt treatment. These interventions can often be time-consuming and costly, and thus providing for all may not be sustainable.

Objective: This study aims to describe our experiences developing a machine learning (ML) model to predict interruption in treatment (IIT) at 30 days among people living with HIV newly enrolled on ART in Nigeria and our integration of the model into the routine information system. In addition, we collected health workers' perceptions and use of the model's outputs for case management.

Methods: Routine program data collected from January 2005 through February 2021 was used to train and test an ML model (boosting tree and Extreme Gradient Boosting) to predict future IIT. Data were randomly sampled using an 80/20 split into training and test data sets, respectively. Model performance was estimated using sensitivity, specificity, and positive and negative predictive values. Variables considered to be highly associated with treatment interruption were preselected by a group of HIV prevention researchers, program experts, and biostatisticians for inclusion in the model. Individuals were defined as having IIT if they were provided a 30-day supply of antiretrovirals but did not return for a refill within 28 days of their scheduled follow-up visit date. Outputs from the ML model were shared weekly with health care workers at selected facilities.

Results: After data cleaning, complete data for 136,747 clients were used for the analysis. The percentage of IIT cases decreased from 58.6% (36,663/61,864) before 2017 to 14.2% (3690/28,046) from October 2019 through February 2021. Overall IIT was higher among clients who were sicker at enrollment. Other factors that were significantly associated with IIT included pregnancy and breastfeeding status and facility characteristics (location, service level, and service type). Several models were initially developed; the selected model had a sensitivity of 81%, specificity of 88%, positive predictive value of 83%, and negative predictive value of 87%, and was successfully integrated into the national electronic medical records database. During field-testing, the majority of users reported that an IIT prediction tool could lead to proactive steps for preventing IIT and improving patient outcomes.

Conclusions: High-performing ML models to identify patients with HIV at risk of IIT can be developed using routinely collected service delivery data and integrated into routine health management information systems. Machine learning can improve the

targeting of interventions through differentiated models of care before patients interrupt treatment, resulting in increased cost-effectiveness and improved patient outcomes.

(*JMIR AI* 2023;2:e44432) doi:[10.2196/44432](https://doi.org/10.2196/44432)

KEYWORDS

HIV; machine learning; treatment interruption; Nigeria; chronic disease; antiretroviral therapy; chronic disease; HIV program; intervention; data collection

Introduction

Antiretroviral therapy (ART) for HIV treatment has transformed HIV from a fatal illness to a lifelong, yet manageable, chronic disease [1]. Long-term adherence to ART and subsequent viral load suppression decrease morbidity and mortality, and reduce the risk of viral transmission [2]. As increasing numbers of countries meet the United Nations Joint Programme on HIV/AIDS (UNAIDS) 95-95-95 benchmarks, tailored interventions and data systems are needed to proactively identify the individuals at highest risk and reduce interruption in treatment (IIT) to achieve and sustain epidemic control [3]. Such data and systems must reflect the reality that retention is not a linear pathway; instead, patients cycle in and out of care. Data from the US President's Emergency Plan for AIDS Relief (PEPFAR) for the period from January 1 to March 31, 2022, show that approximately 4.8% of all patients on ART cycle in and out of treatment (US President's Emergency Plan for AIDS Relief, unpublished data, March 2023). Historically, data from sub-Saharan Africa have suggested that the proportion of individuals remaining on HIV therapy after 3 years has been about 65% [4].

HIV programs use a range of programmatic approaches to support individuals in sustaining adherence to ART and re-engaging those who interrupt treatment [5]. These interventions for preventing IIT or re-engaging those who have already interrupted their treatment can be time-consuming and costly if not targeted. This can lead to inefficiencies from public health, resource management, and sustainability perspectives [6,7]. Innovative approaches to identifying individuals at high risk of IIT and tailored activities to prevent IIT are needed to ensure optimal client health and sustained epidemic control [8,9]. Applying machine learning (ML) for predicting individuals at high risk of IIT paves the way for differentiated service delivery solutions that are individualized, evidence-based, and responsive to improve retention in care and treatment in the path toward epidemic control.

Large data sets containing individual-level data for people living with HIV are now widely available and may create new opportunities to identify patterns and relationships between individual factors and observed client outcomes. Mathematical models can take the process a step further and use retrospective data to predict future behavior [10]. This application of ML is part of a broader trend leveraging artificial intelligence across a range of development sectors, including agriculture, health, and natural disaster response systems [11,12]. HIV use cases have been developed to understand how predictive analytics can improve client services and reduce service delivery pain points across the HIV continuum of care. These use cases

enhance our understanding of the theory of change for how predictive analytics can improve HIV clinical outcomes, program efficiency, and cost-effectiveness. One of the use cases developed in South Africa, termed the "Fall-Out Forecaster," models how recognizing client risk factors can lead to optimized treatment support interventions and minimize IIT. This model could reduce IIT by 6%-10% and reduce care and support costs by 4%-5% in the first 12 months [13].

The real-world application of theoretical HIV use cases of ML in low- and middle-income settings is growing. In Nigeria and Kenya, ML was applied to retrospective patient-level data sets. The models identified independent predictors of IIT among patients receiving ART in Kenya and helped create behavioral risk profiles [14]. In South Africa, retrospective data for clinical, laboratory, and visit patterns were used to develop an ML algorithm that identifies individuals at risk of unsuppressed viral load at their next visit [15]. In Haiti, health care workers used an ML algorithm to generate client risk scores that classified clients into five categories of risk for treatment failure [16]. Health care workers were subsequently trained to provide culturally sensitive, tailored psychosocial counseling to promote retention among clients assessed as high-risk. In South Africa, an ML model helped to define a unique set of retention services tailored for each client [17,18]. In Mozambique, efforts starting in 2018 used ML models to generate risk scores for client likelihood of interrupting treatment (integrated into service delivery via a mobile app or an OpenMRS "plug-in"); the integration demonstrated the ability to rank clients by overall risk, but the ability to plan treatment retention services according to risk level is still under study [19].

In this paper, we describe the development of an ML model to predict IIT at 30 days among people living with HIV newly enrolled in ART in Nigeria and our experiences integrating the model into the routine HIV treatment program. We report the process of model development, early experiences integrating the model into a routine health management information system, and ML users' perceptions and use of the model outputs for case management.

Methods

Program Description

The Strengthening Integrated Delivery of HIV/AIDS Services (SIDHAS) project in Nigeria supports the government of Nigeria in implementing comprehensive HIV services in Akwa Ibom and Cross River states. The goal is to sustain the integration of HIV and AIDS services with tuberculosis (TB) services by building the capacity of the government of Nigeria staff to deliver high-quality, comprehensive, preventive care and

treatment and other related services. The project, which began in 2011, currently supports treatment at 154 health facilities including public, private for-profit, and faith-based organizations; 103 community pharmacies; and 2684 other community ART refill structures. To support case management, individual-level client data are recorded in the electronic medical record system, Lafiya Management Information System (LAMIS).

Data Collection and Cleaning

For this study, we used routine program data from the SIDHAS project to quantify the association of individual characteristics with IIT among people living with HIV receiving ART and developed an ML model to predict future IIT. Data from the patient, clinic, and pharmacy data sets from Akwa Ibom and Cross River states in Nigeria collected from January 2005 through February 2021 were extracted from LAMIS and used for model development. These service delivery data are collected using standardized paper-based forms at each patient encounter and then entered into LAMIS by facility staff. All personal identifiers were removed, and patient data were linked to create one consolidated data set using the unique treatment identification number. We included all patients who were newly enrolled on ART and provided a 30-day supply of antiretrovirals (ARVs) at one of the SIDHAS-supported treatment facilities. The three separate databases were reviewed, and data for selected variables were extracted for all eligible individuals. For the purposes of the study, individuals were defined as having IIT if they were provided a 30-day supply of ARVs but did not return for a refill within 28 days of their scheduled follow-up visit date.

The consolidated data set was subjected to a series of internal consistency checks during which records with invalid data were removed. Reasons for record removal included that the ART start date was listed as earlier than the date of the confirmed HIV test, participants were enrolled too recently to have an observed end point, and the date of the next appointment after enrollment was missing. Participants who were transferred in from other facilities were also excluded given that the interest was in IIT after ART initiation.

Missing data were then addressed for the remaining records in the cleaned data set. Two approaches were used to handle missing data based on the nature of the data collection and operation in the program field. First, missing data within the patient data set were imputed using the k-nearest neighbor algorithm [20] in which the missing value was classified by a plurality vote of its neighbors and the class most common among its k-nearest neighbors was assigned. Second, for variables such as TB status that could not be imputed, missing data within the clinic data set were classified as “missing” in the final data set. In addition, variables such as pregnancy/breastfeeding status for male clients or female clients younger than 10 years or older than 60 years that had incorrect values were categorized as “not applicable.”

Variable Selection

The predictor variables that were used for model building were extracted from the routine health information system. They were

preselected as they were considered to be strongly associated with treatment interruption by a group of SIDHAS project staff and HIV prevention and treatment experts in consultation with biostatisticians. The variables selected for the model included age, gender, marital status, occupation, education, local government area, baseline clinic stage, TB status, pregnancy and breastfeeding status, and facility characteristics (service level, facility type, ownership, population setting, state, ward, and care entry point). The feature (predictor) importance was applied to understand the data and to improve model building and interpretability.

Model Development, Validation, and Testing

The final cleaned data set was randomly divided into a training data set containing 80% of the clients and a test data set with the remaining 20% of the clients. The first data set was used to train predictive models using the 10-fold cross-validation approach, while the second was used to validate model performance. Boosting classification algorithms (eg, boosting tree and Extreme Gradient Boosting) were applied to build predictive models. Positive predictive value, negative predictive value, and Cohen kappa were used to assess the performance of predictive models. The models were further validated on a second data set containing 1107 clients who initiated ART from March through October 2021.

Field Implementation and User Experience

A total of 10 pilot sites were selected for field-testing of the ML model. These sites included primary, secondary, and tertiary service delivery points with adequate patient volume to ensure adequate new client enrollment. The ML algorithm was programmed into LAMIS such that after data from each new patient were entered into the database, the person's IIT chance was automatically generated. At the end of each week, a list that showed the risk of IIT among those provided with a 30-day supply of ARVs was generated and shared with facility staff. Project staff, health care workers, and treatment supporters at the 10 selected facilities were trained on the basics of ML and on the interpretation and application of IIT scores in patient management. Persons with an IIT score of 50% or more were considered to be at high risk for IIT and their case managers provided additional monitoring and assigned an expert to provide psychosocial support through virtual or physical mechanisms to ensure that the client was mentally prepared for the challenges of lifelong ART. All other persons received the standard case management support that is provided to all clients.

Feedback from the health care workers at the pilot sites was collected in two ways. First, we routinely gathered verbal feedback as part of “daily situation room meetings.” These standing meetings were designed to review routine data and gave health care workers a platform to ask questions about the scores, clarify how the tool was working, and contribute practical suggestions for improvement. Second, we collected user feedback formally using a Google Forms questionnaire. The questionnaire in Google Forms was distributed electronically to health care workers at the selected pilot facilities, and they provided written feedback. The form collected information on the sociodemographic characteristics of the respondents; usefulness, acceptance, and relevance of the

ML outputs for improving patient care; experiences interpreting and using the ML scores; and any suggestions for improving the presentation of the scores. The data from the two sources were combined and summarized according to key themes.

Ethical Considerations

The data for this study were collected from an existing project database that is used for routine patient management and program monitoring. The study was reviewed by the Protection of Human Subjects Committee at FHI 360 and was categorized as research not involving human subjects. The authors had no access to patients or personally identifiable information for the individuals whose data were included in the study.

Results

Model Development

After data cleaning, complete data from a total of 136,747 clients were used for the analysis (Figure 1).

The percentage of IIT cases was 41.5% (56,581/136,747) overall but changed over time (Table 1). It decreased significantly during successive years, ranging from 58.6% (36,663/61,864) before 2017 to 14.2% (3690/28,046) during October 2019 through February 2021. Clients sicker at enrollment had higher

IIT rates; IIT was 31.7% (20,465/64,508) among individuals with stage I disease at enrollment compared to 43.5% (12,867/29,557) among those with stage II disease and 59% (2125/3600) among those with stage IV disease. A greater proportion of clients whose baseline clinical stage or baseline clinic data (TB, pregnancy, and breastfeeding status) were missing were classified as IIT compared to individuals with data available for these variables. Other variables that were significantly associated with IIT rates were facility characteristics: location, service level, and service type. IIT rates did not vary significantly by age, gender, education level, marital status, or occupation.

To incorporate the features of the variables, eight models were trained using training data sets with and without year of ART initiation, clinic data (TB, pregnancy, and breastfeeding status), or facility characteristics. The results indicated that models without clinic data would lose more than 10% of predictive accuracy compared to those models with clinic data included, whereas the facility information and year of ART initiation variables only had a slight impact on model performance (Table 2). The results of the model testing on the data from March through October 2021 were similar to the results observed from the test data. These findings indicated that the predictive models were robust and useful for future IIT prediction in the same setting of ART programs.

Figure 1. Study cohort flow diagram. ART: antiretroviral therapy; ARV: antiretroviral; IIT: interruption in treatment.

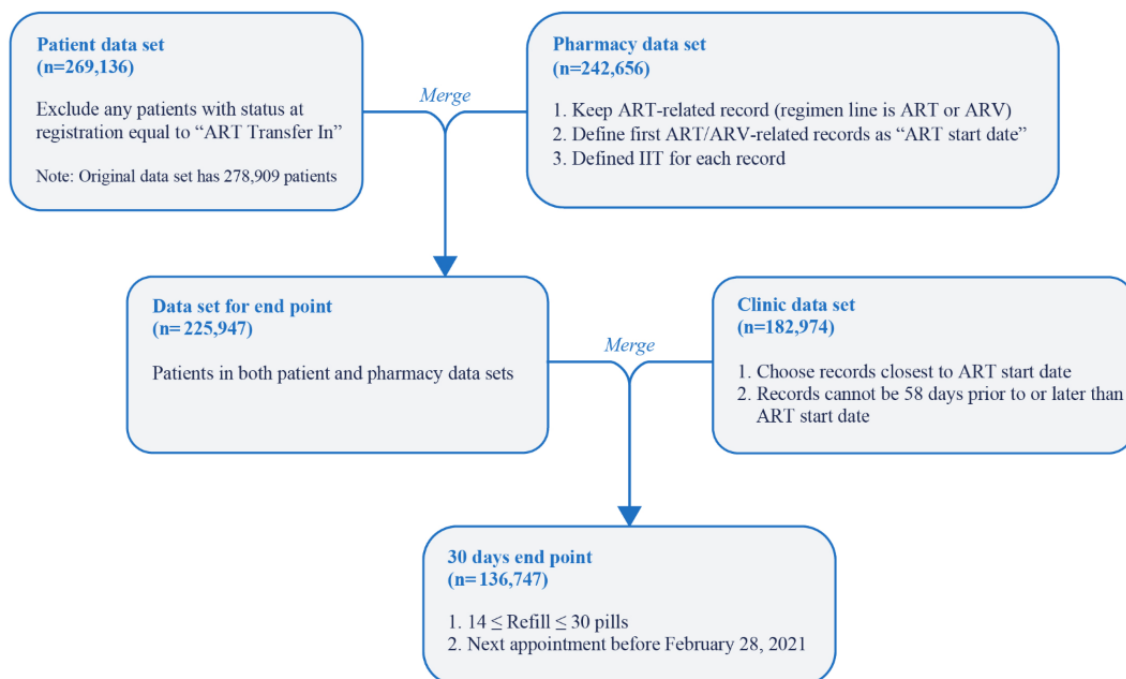


Table 1. Characteristics of the individuals included in the data set used for the model development.

Variable and category	Individuals (N=136,747), n (%)
Interruption in treatment	
Yes	56,581 (41.38)
No	80,166 (58.62)
Year of antiretroviral initiation	
Before 2017	61,939 (45.3)
January 2017-September 2019	46,776 (34.2)
October 2019-February 2021	28,032 (20.5)
Gender	
Female	91,982 (67.26)
Male	44,765 (32.74)
Age (years)	
<14	5657 (4.14)
14-20	8685 (6.35)
21-35	72,049 (52.69)
>35	50,356 (36.82)
Marital status	
Married	1171 (0.86)
Single	64,899 (47.46)
Previously married	52,934 (38.71)
Education	
Primary and Quranic	16,473 (12.05)
≥1 year of secondary	44,219 (32.34)
None	50,912 (37.23)
Occupation	
Employed	36,863 (26.96)
Unemployed/retired/students	79,059 (57.81)
State	
Akwa Ibom	100,937 (73.82)
Cross River	35,791 (26.18)
Baseline clinic stage	
Stage I	64,508 (47.17)
Stage II-IV	68,740 (50.27)
Facility type	
Health center/clinic/posts	77,597 (56.75)
General, tertiary, or cottage hospital	59,131 (43.25)
TB^a status^b	
No signs or symptoms of TB	58,953 (43.11)
Currently on isoniazid prophylaxis	4745 (3.47)
Confirmed/suspected TB	5167 (3.8)
Pregnant^c	
No	46,883 (51.0)

Variable and category	Individuals (N=136,747), n (%)
Yes	925 (1.0)
Breastfeeding^c	
No	47,617 (51.8)
Yes	191 (0.21)

^aTB: tuberculosis.

^bTotals do not add up to 136,747 for all variables under TB status due to missing values for some variables.

^cn=91,982 (number of females in the data set).

Table 2. Model performance evaluation with test data from January 2005 through February 2021 and validation data for model 4.

	Model 1 ^a	Model 2 ^b	Model 3 ^c	Model 4 ^d (selected) ^e	Model 4 ^f	Model 5 ^g	Model 6 ^h	Model 7 ⁱ	Model 8 ^j
Accuracy (95% CI)	0.85 (0.85-0.86)	0.83 (0.83-0.84)	0.87 (0.87-0.87)	0.85 (0.85-0.86)	0.91 (0.88-0.93)	0.75 (0.74-0.75)	0.70 (0.69-0.70)	0.75 (0.74-0.75)	0.72 (0.72-0.73)
Sensitivity (95% CI)	0.82 (0.81-0.82)	0.75 (0.75-0.76)	0.84 (0.83-0.84)	0.81 (0.81-0.82)	0.79 (0.73-0.86)	0.63 (0.62-0.64)	0.58 (0.57-0.59)	0.63 (0.62-0.64)	0.62 (0.61-0.63)
Specificity (95% CI)	0.88 (0.87-0.88)	0.89 (0.88-0.89)	0.89 (0.89-0.90)	0.88 (0.88-0.89)	0.94 (0.92-0.96)	0.83 (0.82-0.84)	0.78 (0.77-0.78)	0.83 (0.82-0.84)	0.80 (0.79-0.81)
PPV ^k (95% CI)	0.83 (0.82-0.83)	0.82 (0.82-0.83)	0.85 (0.84-0.85)	0.83 (0.82-0.83)	0.77 (0.70-0.83)	0.72 (0.72-0.73)	0.65 (0.64-0.66)	0.72 (0.71-0.73)	0.69 (0.68-0.70)
NPV ^l (95% CI)	0.87 (0.87-0.88)	0.84 (0.83-0.84)	0.89 (0.88-0.89)	0.87 (0.87-0.88)	0.94 (0.93-0.96)	0.76 (0.76-0.77)	0.72 (0.72-0.73)	0.76 (0.76-0.77)	0.75 (0.74-0.75)
Kappa	0.69	0.65	0.73	0.70	0.72	0.47	0.36	0.47	0.42

^aModel 1 included clinic variables (tuberculosis [TB], pregnancy, and breastfeeding status) and year of antiretroviral therapy (ART) initiation.

^bModel 2 included clinic variables (TB, pregnancy, and breastfeeding status).

^cModel 3 included clinic variables (TB, pregnancy, and breastfeeding status), facility information, and year of ART initiation.

^dModel 4 included clinic variables (TB, pregnancy, and breastfeeding status) and facility information.

^eModel selected for application.

^fValidation data March-November 2021.

^gModel 5 included year of ART initiation and did not include clinical variables (TB, pregnancy, and breastfeeding status).

^hModel 6 did not include clinic variables (TB, pregnancy, and breastfeeding status), facility information, and year of ART initiation.

ⁱModel 7 included facility information and year of ART initiation and did not include clinic variables (TB, pregnancy, and breastfeeding status).

^jModel 8 included facility information and did not include clinic variables (TB, pregnancy, and breastfeeding status) and year of ART initiation.

^kPPV: positive predictive value.

^lNPV: negative predictive value.

Field Implementation and User Experience

The 30-day predictive model was integrated into LAMIS and applied to 25 consecutive people living with HIV newly enrolled on ART at selected hospitals and who were provided with a 30-day supply of an ART regimen over a 15-week period (April to July 2022). None were seen to be a high risk for IIT based on the predetermined 50% threshold. The predicted IIT risks ranged from 1.8% to 25.7%. All clients received routine psychosocial support, monitoring of possible adverse drug reactions, and overall support through virtual check-ins and home visits. Given that their risk prediction scores did not meet the 50% threshold, additional intensive services were not provided. Changes in local policies promoting multi-month dispensing of ARVs to people living with HIV have resulted in the majority of those who are newly enrolled on ART being provided with a 90-day supply of medication and a smaller proportion provided with a 30-day supply of ARVs.

Of the 48 individuals who provided feedback on usability and acceptability, 36 (75%) indicated that the IIT prediction tool was useful. Common reasons they cited included early notification to the site of a client with high IIT potential and the ability to improve case management at the site, thus helping patient management and monitoring be more proactive than reactive. As one facility backstop mentioned:

It has helped us to monitor our clients, calling them up and giving them a timeline to come for their refills so that their treatment won't be interrupted.

While most data entry clerks and monitoring and evaluation specialists provided positive feedback on accessibility, a few were skeptical or neutral. Those with a positive view indicated that since the model was integrated into LAMIS, Nigeria's routine national HIV information system, rather than a secondary application, it was straightforward and easy to navigate. One data entry clerk reported:

My experience using the machine learning predictive is that as a data entry clerk I will use the machine to check and relate with my case manager to track the client in time to avoid IIT.

A monitoring and evaluation specialist from a primary health center said:

At first, I found it challenging to understand the chance of IIT, but after understanding and using it, I now see it as indices to protect our program growth from negative adjustment.

One of the more skeptical data entry clerks related that:

I haven't seen to understand the logic behind it...The outcome didn't change the restart or return to care. I need the ideas behind this...

Discussion

Principal Results

Using routinely collected service delivery data, we developed an ML model to predict IIT among people living with HIV in Nigeria that was easy to introduce and acceptable to providers in routine clinical care settings. All models developed included the use of routinely collected individual- and clinic-level variables to determine the risk of IIT among clients receiving a 30-day supply of ART. The final model chosen had both sensitivity and positive predictive values higher than 80%. After initial challenges, our model was successfully incorporated into the national systems for routine individual-level case management and monitoring and evaluation in pilot clinics. We found health care workers to be amenable to incorporating the prediction tool into routine work and eager to increase opportunities to tailor interventions to those most in need. Our ML model performed well on our test data and integrated well into routine systems but has yet to be deployed and assessed for effectiveness at the population level.

Limitations

The low number of clients receiving 30 days of ART limited our ability to make programmatic adjustments based on the likelihood of IIT and prevented the prospective assessment of performance or effectiveness. As multi-month scripting is now the norm, models incorporating the multi-month dosing data or developing a new model to be used among clients receiving 3 months or more of ART are needed. Additionally, more work is needed to understand the sensitivity and specificity of the model on IIT after the first 30 days and the usefulness of these models outside of the population or geography on which they were based.

The limitations that are inherent in routinely collected service delivery data will also need to be addressed before these data are used for developing ML models. In Nigeria, as in many countries, social and contextual community factors were not routinely collected in their national health management information system and thus were not factored into the model despite known associations with IIT. In our data set, we encountered high levels of missing and misclassified data that were handled statistically yet are illustrative of the challenges

related to data quality. After the incorporation of the model into LAMIS, staff took greater care to address delayed and incomplete data entry, resulting in a significant reduction in the proportion of missing data. HIV programs have changed over time and continue to change quickly. Developing a model based on retrospective data is a limitation, and models must be tested prospectively to determine if the accuracy holds with newer data. As the wealth of programmatic data continues to grow, refining models as a tool to target services and improve the quality of care will be critical.

Comparison With Prior Work and Implications

Using ML to improve continuity of HIV care is a practical example of how advanced analytics can address population- and individual-level global health challenges, as we continue to advance digital health maturity [14,21]. While ML analytics hold great promise for closing the final gaps to achieve the 95-95-95 targets, the representativeness of available and accessible data must be considered [15,22]. With representative data, ML models enable us to limit biases and increase service equity based on standard algorithms. In addition, ML models could be a useful tool that future programs could use to tailor interventions to a person's unique needs. This can decrease differences in the quality of health care across sites or decrease the perpetuation of any health care worker bias against some vulnerable populations. From a sustainability perspective, addressing constraints in digital infrastructure and human resources are critical investments for scaling country-owned predictive analytics for addressing IIT and other important public health issues. Investments in this area can also contribute to the growth of a country's broader digital health system architecture.

Recently, there have been increasing efforts in low- and middle-income settings to develop and integrate predictive analytics into public health programs and to demonstrate that these tools can be implemented in low-resource settings. In any setting where optimization is critical due to labor or fiscal shortages, ML can help target the efficient use of human and financial resources. However, it is critical to consider broader partnerships, deployment, and scaling of ML to ensure that ongoing investments are strategic and sustainable. Such solutions may require additional budgeting for foundational infrastructure (eg, connectivity, cybersecurity, cloud housing, data management, electricity access), along with the human resources and capacity building needed for ongoing independent program support. Determining when it is strategic to invest in ML given the broader investments required for sustainable ML and the wide range of HIV interventions available to improve treatment continuity will require assessing cost-effectiveness. Considering costing and evaluation methodologies and prioritizing investments that benefit the strengthening of broader digital infrastructure are opportunities to realize economies of scale and a greater return on investment.

Conclusions and Next Steps

Despite initial challenges, we were able to successfully develop and deploy an ML model into LAMIS, Nigeria's routine HIV information system. There was a high level of acceptance of the ML model among staff at the pilot facilities. Our model will

be refined as additional data are made available; this includes expansion to include IIT in the context of multi-month dosing. The model will be assessed with prospective data to refine the appropriate cutoff for determining high risk and thus the threshold for providing additional services.

Acknowledgments

The authors wish to acknowledge the contributions of everyone who was involved in the Strengthening Integrated Delivery of HIV/AIDS Services (SIDHAS) project in Nigeria, particularly the technical and strategic information staff based at the various facilities.

This work was made possible by the generous support of the American people through the US Agency for International Development (USAID) and the US President's Emergency Plan for AIDS Relief (PEPFAR). This publication features data collected during the implementation of the PEPFAR-funded SIDHAS project in Nigeria (#AID-620-A-11-00002). The data analysis and preparation of the manuscript were funded by FHI 360 and PEPFAR through the Meeting Targets and Maintaining Epidemic Control (EpiC) Project (#7200AA19CA00002). The contents are the responsibility of the authors and do not necessarily reflect the views of USAID, PEPFAR, or the US Government.

Authors' Contributions

NEP, MDO, CFW, AAG, AI, and EH conceptualized the paper. MDO, AI, MTL, PLC, JS, and AAG were involved in the literature review, data compilation, and analysis. All authors contributed to data interpretation and manuscript drafts and approved the final version.

Conflicts of Interest

None declared.

References

1. Deeks SG, Lewin SR, Havlir DV. The end of AIDS: HIV infection as a chronic disease. *Lancet* 2013 Nov 02;382(9903):1525-1533 [FREE Full text] [doi: [10.1016/S0140-6736\(13\)61809-7](https://doi.org/10.1016/S0140-6736(13)61809-7)] [Medline: [24152939](https://pubmed.ncbi.nlm.nih.gov/24152939/)]
2. Eisinger RW, Dieffenbach CW, Fauci AS. HIV viral load and transmissibility of HIV infection: undetectable equals untransmittable. *JAMA* 2019 Feb 05;321(5):451-452. [doi: [10.1001/jama.2018.21167](https://doi.org/10.1001/jama.2018.21167)] [Medline: [30629090](https://pubmed.ncbi.nlm.nih.gov/30629090/)]
3. PEPFAR 2022 Country and Regional Operational Plan (COP/ROP) guidance for all PEPFAR-supported countries. US Department of State. URL: https://www.state.gov/wp-content/uploads/2022/02/COP22-Guidance-Final_508-Compliant-3.pdf [accessed 2023-04-20]
4. Kranzer K, Govindasamy D, Ford N, Johnston V, Lawn SD. Quantifying and addressing losses along the continuum of care for people living with HIV infection in sub-Saharan Africa: a systematic review. *J Int AIDS Soc* 2012 Nov 19;15(2):17383 [FREE Full text] [doi: [10.7448/IAS.15.2.17383](https://doi.org/10.7448/IAS.15.2.17383)] [Medline: [23199799](https://pubmed.ncbi.nlm.nih.gov/23199799/)]
5. Rajabiun S, Mallinson RK, McCoy K, Coleman S, Drainoni M, Rebholz C, et al. "Getting me back on track": the role of outreach interventions in engaging and retaining people living with HIV/AIDS in medical care. *AIDS Patient Care STDS* 2007;21 Suppl 1:S20-S29. [doi: [10.1089/apc.2007.9990](https://doi.org/10.1089/apc.2007.9990)] [Medline: [17563286](https://pubmed.ncbi.nlm.nih.gov/17563286/)]
6. Mirzazadeh A, Eshun-Wilson I, Thompson RR, Bonyani A, Kahn JG, Baral SD, et al. Interventions to reengage people living with HIV who are lost to follow-up from HIV treatment programs: a systematic review and meta-analysis. *PLoS Med* 2022 Mar;19(3):e1003940 [FREE Full text] [doi: [10.1371/journal.pmed.1003940](https://doi.org/10.1371/journal.pmed.1003940)] [Medline: [35290369](https://pubmed.ncbi.nlm.nih.gov/35290369/)]
7. Palacio-Vieira J, Reyes-Urueña JM, Imaz A, Bruguera A, Force L, Llaveria AO, PICIS study group. Strategies to reengage patients lost to follow up in HIV care in high income countries, a scoping review. *BMC Public Health* 2021 Aug 28;21(1):1596 [FREE Full text] [doi: [10.1186/s12889-021-11613-y](https://doi.org/10.1186/s12889-021-11613-y)] [Medline: [34454444](https://pubmed.ncbi.nlm.nih.gov/34454444/)]
8. VERSION 2.0 – Draft Overview PEPFAR Strategy: Vision 2025. U.S. Department of State. 2022. URL: https://www.state.gov/wp-content/uploads/2021/09/DRAFT-Overview-PEPFAR-Strategy-Vision-2025_Version-2.0-2.pdf [accessed 2023-04-28]
9. End inequalities. End AIDS. Global AIDS strategy 2021-2026. UNAIDS. 2022. URL: <https://www.unaids.org/en/Global-AIDS-Strategy-2021-2026> [accessed 2023-04-27]
10. Olatosi B, Vermund S, Li X. Power of Big Data in ending HIV. *AIDS* 2021 May 01;35(Suppl 1):S1-S5. [doi: [10.1097/QAD.0000000000002888](https://doi.org/10.1097/QAD.0000000000002888)] [Medline: [33867484](https://pubmed.ncbi.nlm.nih.gov/33867484/)]
11. Reflecting the past, shaping the future: making AI work for international development. U.S. Agency for International Development. URL: <https://www.usaid.gov/digital-development/machine-learning/AI-ML-in-development> [accessed 2023-04-27]
12. Managing machine learning projects in international development: a practical guide. U.S. Agency for International Development. 2021. URL: <https://www.usaid.gov/digital-development/managing-machine-learning-projects> [accessed 2023-04-27]

13. Data and advanced analytics in HIV service delivery: use cases to help reach 95-95-95. U.S. Agency for International Development. 2020. URL: <https://www.usaid.gov/cii/data-advanced-analytics> [accessed 2023-04-27]
14. Wang B, Liu F, Deveaux L, Ash A, Gosh S, Li X, et al. Adolescent HIV-related behavioural prediction using machine learning: a foundation for precision HIV prevention. *AIDS* 2021 May 01;35(Suppl 1):S75-S84 [FREE Full text] [doi: [10.1097/QAD.0000000000002867](https://doi.org/10.1097/QAD.0000000000002867)] [Medline: [33867490](https://pubmed.ncbi.nlm.nih.gov/33867490/)]
15. Ashrafian H, Darzi A. Transforming health policy through machine learning. *PLoS Med* 2018 Nov;15(11):e1002692 [FREE Full text] [doi: [10.1371/journal.pmed.1002692](https://doi.org/10.1371/journal.pmed.1002692)] [Medline: [30422977](https://pubmed.ncbi.nlm.nih.gov/30422977/)]
16. Puttkammer N, Simoni JM, Sandifer T, Chéry JM, Dervis W, Balan JG, et al. An EMR-based alert with brief provider-led ART adherence counseling: promising results of the InfoPlus Adherence pilot study among Haitian adults with HIV initiating ART. *AIDS Behav* 2020 Dec;24(12):3320-3336 [FREE Full text] [doi: [10.1007/s10461-020-02945-8](https://doi.org/10.1007/s10461-020-02945-8)] [Medline: [32715409](https://pubmed.ncbi.nlm.nih.gov/32715409/)]
17. Maskew M, Sharpey-Schafer K, De Voux L, Bor J, Rennick M, Crompton T, et al. Machine learning to predict retention and viral suppression in South African HIV treatment cohorts. *ReadCube*. 2019. URL: <https://www.readcube.com/articles/10.1101%2F2021.02.03.21251100> [accessed 2023-04-27]
18. Upchurch K. 14 September 2021 DUC Meeting 11 summary "Proactive Adherence Counseling". OpenHIE Wiki. 2021 Sep 14. URL: <https://wiki.ohie.org/pages/viewpage.action?pageId=83398041> [accessed 2023-04-27]
19. Machine learning for predicting default from HIV services in Mozambique: OpCon Mozambique final report. ICAP. 2019. URL: https://icap.columbia.edu/wp-content/uploads/OpCon-Mozambique_Final-Report_FINAL.pdf [accessed 2023-04-27]
20. Beretta L, Santaniello A. Nearest neighbor imputation algorithms: a critical evaluation. *BMC Med Inform Decis Mak* 2016 Jul 25;16 Suppl 3(Suppl 3):74 [FREE Full text] [doi: [10.1186/s12911-016-0318-z](https://doi.org/10.1186/s12911-016-0318-z)] [Medline: [27454392](https://pubmed.ncbi.nlm.nih.gov/27454392/)]
21. Stockman J, Friedman J, Sundberg J, Harris E, Bailey L. Predictive analytics using machine learning to identify ART clients at health system level at greatest risk of treatment interruption in Mozambique and Nigeria. *J Acquir Immune Defic Syndr* 2022 Jun 01;90(2):154-160. [doi: [10.1097/QAI.0000000000002947](https://doi.org/10.1097/QAI.0000000000002947)] [Medline: [35262514](https://pubmed.ncbi.nlm.nih.gov/35262514/)]
22. Sambasivan N, Arnesen E, Hutchinson B, Doshi T, Prabhakaran V. Re-imagining algorithmic fairness in India and beyond. In: *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*. 2021 Mar Presented at: FAccT '21; March 3-10, 2021; Virtual Event, Canada p. 315-328. [doi: [10.1145/3442188.3445896](https://doi.org/10.1145/3442188.3445896)]

Abbreviations

ART: antiretroviral therapy

ARV: antiretroviral

IIT: interruption in treatment

LAMIS: Lafiya Management Information System

ML: machine learning

PEPFAR: US President's Emergency Plan for AIDS Relief

SIDHAS: Strengthening Integrated Delivery of HIV/AIDS Services

TB: tuberculosis

UNAIDS: United Nations Joint Programme on HIV/AIDS

Edited by G Luo; submitted 18.11.22; peer-reviewed by P Wang, P Dunn; comments to author 23.02.23; revised version received 16.03.23; accepted 03.04.23; published 12.05.23.

Please cite as:

Ogbechie MD, Fischer Walker C, Lee MT, Abba Gana A, Oduola A, Idemudia A, Edor M, Harris EL, Stephens J, Gao X, Chen PL, Persaud NE

Predicting Treatment Interruption Among People Living With HIV in Nigeria: Machine Learning Approach

JMIR AI 2023;2:e44432

URL: <https://ai.jmir.org/2023/1/e44432>

doi: [10.2196/44432](https://doi.org/10.2196/44432)

PMID: [38875546](https://pubmed.ncbi.nlm.nih.gov/38875546/)

©Matthew-David Ogbechie, Christa Fischer Walker, Mu-Tien Lee, Amina Abba Gana, Abimbola Oduola, Augustine Idemudia, Matthew Edor, Emily Lark Harris, Jessica Stephens, Xiaoming Gao, Pai-Lien Chen, Navindra Etwaroo Persaud. Originally published in *JMIR AI* (<https://ai.jmir.org>), 12.05.2023. This is an open-access article distributed under the terms of the Creative Commons Attribution License (<https://creativecommons.org/licenses/by/4.0/>), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in *JMIR AI*, is properly cited. The complete bibliographic information, a link to the original publication on <https://www.ai.jmir.org/>, as well as this copyright and license information must be included.

Original Paper

Machine Learning–Based Time in Patterns for Blood Glucose Fluctuation Pattern Recognition in Type 1 Diabetes Management: Development and Validation Study

Nicholas Berin Chan¹, MScR; Weizi Li¹, DPhil; Theingi Aung², MBBS, MD; Eghosa Bazuaye², MSc; Rosa M Montero³, MRCP, MBBS, MD

¹Informatics Research Centre, Henley Business School, University of Reading, Reading, United Kingdom

²Royal Berkshire NHS Foundation Trust, Reading, United Kingdom

³King's College London, London, United Kingdom

Corresponding Author:

Weizi Li, DPhil

Informatics Research Centre

Henley Business School

University of Reading

Whiteknights

Reading, RG6 6UD

United Kingdom

Phone: 44 7714021891

Email: weizi.li@henley.ac.uk

Abstract

Background: Continuous glucose monitoring (CGM) for diabetes combines noninvasive glucose biosensors, continuous monitoring, cloud computing, and analytics to connect and simulate a hospital setting in a person's home. CGM systems inspired analytics methods to measure glycemic variability (GV), but existing GV analytics methods disregard glucose trends and patterns; hence, they fail to capture entire temporal patterns and do not provide granular insights about glucose fluctuations.

Objective: This study aimed to propose a machine learning–based framework for blood glucose fluctuation pattern recognition, which enables a more comprehensive representation of GV profiles that could present detailed fluctuation information, be easily understood by clinicians, and provide insights about patient groups based on time in blood fluctuation patterns.

Methods: Overall, 1.5 million measurements from 126 patients in the United Kingdom with type 1 diabetes mellitus (T1DM) were collected, and prevalent blood fluctuation patterns were extracted using dynamic time warping. The patterns were further validated in 225 patients in the United States with T1DM. Hierarchical clustering was then applied on time in patterns to form 4 clusters of patients. Patient groups were compared using statistical analysis.

Results: In total, 6 patterns depicting distinctive glucose levels and trends were identified and validated, based on which 4 GV profiles of patients with T1DM were found. They were significantly different in terms of glycemic statuses such as diabetes duration ($P=.04$), glycated hemoglobin level ($P<.001$), and time in range ($P<.001$) and thus had different management needs.

Conclusions: The proposed method can analytically extract existing blood fluctuation patterns from CGM data. Thus, time in patterns can capture a rich view of patients' GV profile. Its conceptual resemblance with time in range, along with rich blood fluctuation details, makes it more scalable, accessible, and informative to clinicians.

(JMIR AI 2023;2:e45450) doi:[10.2196/45450](https://doi.org/10.2196/45450)

KEYWORDS

diabetes mellitus; continuous glucose monitoring; glycemic variability; glucose fluctuation pattern; temporal clustering; scalable metrics

Introduction

Background

Diabetes mellitus (DM) is a lifelong condition owing to elevated glucose concentration in blood and has long been a major global public health issue. According to the International Diabetes Federation, the number of people with diabetes has risen from 151 million in 2000 to 537 million in 2021 and is projected to reach 783 million by 2045 [1]. The World Health Organization estimated that 1.5 million deaths were directly caused by diabetes in 2019, making it the ninth leading cause of death [2]. Before the introduction of smart and connected health and hence continuous glucose monitoring (CGM) wearable devices, self-monitoring of blood glucose (BG) level played a crucial role in the management of patients with DM. However, a landmark paper in 2008 revealed that patients rarely measured glucose levels after meals or overnight, which led to postprandial hyperglycemia within the group of patients [3]. Results from a multicenter randomized control trial further illustrated that the use of CGM is associated with improved glycemic control in adults with type 1 DM (T1DM). CGM for diabetes combines noninvasive glucose biosensors, continuous monitoring, cloud computing, and analytics to connect and simulate a hospital setting in a person's home. It uses sensors to measure glucose levels just beneath the surface of the skin and sends data wirelessly to the users' compatible smart device or receiver [4]. CGM works as a connected and closed-loop system that enables patients to modify their insulin dosages based on their glucose trends in a timely manner. With the advancement of technology, CGM has become much more accurate and assessable, making it a vital tool for patients with DM to manage their BG level. According to a survey in 2019, the percentage of CGM users with T1DM in the US T1D Exchange registry has increased from 7% in 2010 to 30% in 2018 [5]. A systematic review and meta-analysis in 2019 concluded that the use of CGM over self-monitoring is beneficial in terms of several clinical outcomes [6].

Average BG to Glycemic Variability

As suggested by Huisman et al [7] and characterized by Bookchin and Gallop [8], glycated hemoglobin (HbA_{1c}) level has been the gold standard for testing BG intensity and defining diabetes since its proposal. It is a measure of average glucose within a person over the previous 8 to 12 weeks [9] and has been adopted by major clinical guidelines for managing the glycemic status of patients with T1DM and diagnosing and screening people who are at risk of type 2 DM [10-12].

The introduction of CGM opened up new areas of research for BG control owing to the sheer volume of BG data it collects. Despite the well-recognized evidence and wide use of HbA_{1c} level, there has been increasing research interest in glycemic variability (GV), which is based on CGM data, arguing that GV contains additional diagnostic and prognostic value that could not be fully captured by HbA_{1c} measurement. BG variability, also known as GV, refers to the degree of oscillation in BG levels [13]. Patients with diabetes often rely heavily on continuous medication intake to maintain BG at a normal and stable level. However, this is often difficult as food consumption

would lead to a spike in BG, whereas the use of excessively intensive medication could lead to hypoglycemia. As HbA_{1c} measurement fails to effectively capture these oscillations, HbA_{1c} level alone is not an ideal indicator of an individual patient's glycemic control [14]. Studies have been conducted to evaluate the diagnostic and prognostic value of GV. It is shown that high GV is associated with high risk of microvascular and macrovascular complications [15,16], high mortality in patients who are critically ill [17-19], and high incidence of neurological outcomes [20]. A systematic review and meta-analysis conducted by Gorst et al [21] indicated that high GV is associated with increased risk of renal disease, cardiovascular events, retinopathy, ulceration, and mortality.

Quantifying GV

Several methods have been proposed to capture GV from CGM data. SD and coefficient of variation (COV) are the 2 most prevalent metrics in the field owing to their ease of calculation and relative understandability. However, they are often criticized as a statistically biased metric to represent GV because BG readings do not follow a normal distribution and tend to skew toward hyperglycemia, especially in patients with diabetes [22,23]. In addition, they do not incorporate the information about time and sequences of readings in their calculations. As such, even if one randomly reorders a set of BG readings to obtain drastically different glycemic curves, the SD and COV would still remain the same.

Time in range (TIR) has been proposed by existing studies as a way to indirectly capture GV [24-29]. TIR refers to the daily proportion of time one's glucose level falls within given target ranges with breakpoints typically at 3, 3.9, 10, and 13.9 mmol/L [29]. The major strengths of TIR are that it can be readily computed and it is much more intuitive to clinicians, while still, to some extent, able to capture how much a person's BG deviates from the target range. So far, studies have shown that TIR alone is associated with a wide range of outcomes, such as diabetic retinopathy [26] and various neonatal outcomes [30]. A conference conducted in 2018 reached a consensus that outlined the use of CGM and related glycemic metrics to improve glucose management [27,28]. Despite the widely recognized strengths of TIR, its aggregated nature inevitably implies that temporal fluctuation information from CGM data is left unused, which was shown to contain further prognostic value. In particular, as TIR also disregards the order in which the glucose measurements were made, it fails to provide details about specific glycemic patterns that occurred in one's CGM history.

Most metrics fail to account for the sequences of BG measurements without the use of sophisticated statistical or machine learning models because that would involve recognizing a trend or pattern within a time series of BG data. Thus, machine learning models have also been proposed to compute GV. Struble [31] and Marling et al [32] applied support vector regression to model the data points from CGM and computed GV based on the difference between actual and modeled data points. Eljil et al [33] suggested the use of time-sensitive artificial neural networks to predict hypoglycemic events, whereas Mani et al [34] used random forest models to

predict the risk of type 2 DM. Furthermore, Hall et al [35] defined 3 glucose fluctuation patterns, namely low, medium, and high variability, by using dynamic time warping (DTW). A list of analytic methods and metrics for quantifying GV in existing literature is summarized in Table 1. Although these

machine learning–based methods successfully used the temporal information embedded in CGM data, they were criticized to be “not well understood in clinical practice” [36], which remains as a major hurdle that hinders clinicians from applying these methods in practice.

Table 1. Summary of metrics and analytics methods for assessing glycemic variability (GV).

Metrics and analytic methods	Related publications	Strengths	Limitations
SD	Krinsley [19]	Simplicity	<ul style="list-style-type: none"> • Tend to be skewed and does not adjust for mean BG^a level • Does not account for temporal information • Limited capability of interpreting GV profiles with only a single value
COV ^b	Rodbard [37] and Rama Chandran et al [38]	Simplicity and adjusts for mean	<ul style="list-style-type: none"> • Does not account for temporal information • Limited capability of interpreting GV profiles with only a single value
TIR ^c	Omar et al [24], Beck et al [25], Lu et al [26], Beyond A1C Writing Group [27], Battelino et al [28], and Advani [29]	Simplicity	<ul style="list-style-type: none"> • Does not account for sequence of BG measurements
IQR	McDonnel et al [39]	Simplicity	<ul style="list-style-type: none"> • Does not adjust for mean BG level • Does not account for temporal information • Limited capability of interpreting GV profiles with only a single value
Range	Oh et al [20]	Simplicity	<ul style="list-style-type: none"> • Tend to be skewed and does not adjust for mean BG level • Does not account for temporal information • Limited capability of interpreting GV profiles with only a single value
MAGE ^d	Service [22] and Service et al [40]	Takes BG fluctuation owing to meal into account	<ul style="list-style-type: none"> • Day based • Does not adjust for mean BG level • Does not account for sequences of BG measurements • Limited capability of interpreting GV profiles with only a single value
LBGI ^e and HBGI ^f	Kovatchev et al [41] and Hill et al [42]	Adjusts for BG skewness and measuring frequency	<ul style="list-style-type: none"> • Does not account for sequences of BG measurements • Ambiguities in BG variability level • Limited capability of interpreting GV profiles with only a single value
SVR ^g	Struble [31] and Marling et al [32]	Accounts for temporal information	<ul style="list-style-type: none"> • Limited capability of interpreting GV profiles with only 3 discrete levels • Subject to clinicians' experience in determining the variability levels; thus, lack of evidence
TS-ANN ^h	Eljil et al [33]	Accounts for temporal information	<ul style="list-style-type: none"> • Limited capability of interpreting GV profiles with only a single value
RF ⁱ	Mani et al [34]	Accounts for temporal information	<ul style="list-style-type: none"> • Limited capability of interpreting GV profiles with only a single value
Glucotypes	Hall et al [35]	Accounts for temporal information	<ul style="list-style-type: none"> • Limited capability of interpreting GV profiles with only 3 discrete levels

^aBG: blood glucose.

^bCOV: coefficient of variation.

^cTIR: time in range.

^dMAGE: mean amplitude of glycemic excursions.

^eLBGI: low blood glucose index.

^fHBGI: high blood glucose index.

^gSVR: support vector regression.

^hTS-ANN: time-sensitive artificial neural network.

ⁱRF: random forest.

Furthermore, there has been scalability issues in existing CGM-related machine learning studies owing to the missingness of key variables in real-world application. For example, in most studies, participants are asked to manually log daily events (such as meal, stress level, exercise, and illnesses) and wear a wristband for collecting physiological data, which can potentially provide insights about GV management [43]. However, in real-world health care, most of the time, only routinely collected CGM and electronic patient record (EPR) data would be available for clinicians to make decisions about therapeutic pathways. A more scalable analytical framework is warranted to make full use of CGM data and capture detailed GV pattern to inform personalized therapeutic pathways. Computationally simple methods such as COV and TIR tend to show a narrow presentation of a patient's GV profile but are more recognized among clinicians and used in more clinical studies. In contrast, despite being able to capture more information from CGM data, complex machine learning-based methods tend to be less intuitive for clinicians to apply in practice. Moreover, existing methods often express GV profile as a single value or a few discrete levels (usually high, medium, or low) and do not reveal detailed insights about any GV patterns that exist in the data.

In this study, we sought to address the scalability issues of machine learning-based GV management and fill the gap between the intuitiveness of simplistic methods, such as TIR, and comprehensiveness of machine learning methods to understand the underlying GV patterns in patients with T1DM who have been using wearable CGM. The aim of this paper was 2-fold. First, we sought to develop a novel and scalable analytics framework for efficient GV pattern recognition and attribution that provides a more comprehensive, easy-to-understand representation of a patient's BG fluctuation profile, which cannot

be solely captured by clinically established metrics such as HbA_{1c} level and TIR. Second, we sought to propose the use of time in patterns to depict GV profiles and show that it reveals additional insights about CGM data and patient characteristics. In the long run, we hope that having a rich and accessible representation of GV profile could serve as a step toward explainable artificial intelligence and the development of personalized therapeutic pathways for patients with T1DM.

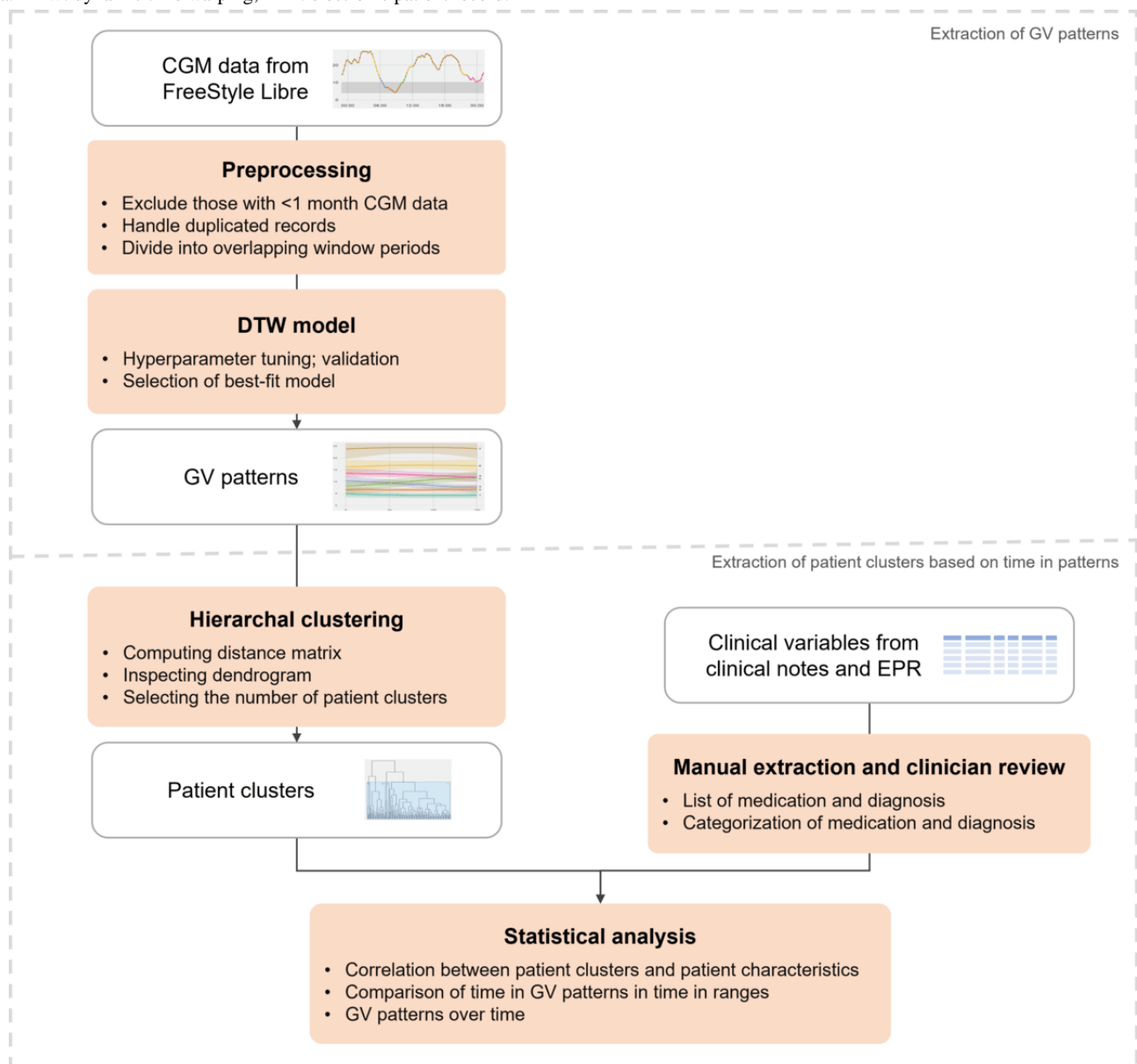
Methods

Overview

The analysis of this study entailed two major parts (Figure 1): (1) extracting GV patterns from CGM and (2) clustering patients based on time in GV patterns and evaluating the clusters. For the first part, we gathered and filtered patients, extracted and cleaned their CGM data from monitoring devices, and then applied a machine learning algorithm called DTW. This enabled us to classify the given CGM data within a time window into one of the extracted patterns. In addition, we applied our methods to another CGM data set to externally validate our pattern extraction methods. In the second part, we computed the time spent in each pattern per patient. Clinical variables were gathered from EPR and clinical notes. Clustering methods were further applied on time in patterns to demonstrate its possible use cases by comparing the differences in clinical variables across clusters of patients. Finally, we evaluated the relationship between time in patterns developed using our method and well-established glycemic metrics.

All analyses were performed using R (version 4.0.3; R Core Team and the R Foundation for Statistical Computing), and the R package *dtwclust* (version 5.5.12, Sarda-Espinosa) was used for DTW-related analyses [44,45].

Figure 1. Analytical framework for glycemic variability (GV) pattern extraction and patient clustering from continuous glucose monitoring (CGM) data. DTW: dynamic time warping; EPR: electronic patient record.



Inclusion and Exclusion of Patients

All patients in this study attended the Centre for Diabetes and Endocrinology of a large hospital in the United Kingdom. The inclusion criteria included patients who (1) were diagnosed with T1DM and (2) were given a CGM device named FreeStyle Libre (FSL) before August 5, 2019, and had been using it for at least one month. Patients aged <18 years or patients with unavailable or missing National Health Service (NHS) identifiers were excluded from this study. Of 130 patients with available CGM data in FSL, 126 (96.9%) patients were included in this study.

Collection of CGM Data

FSL flash glucose monitoring system was used to measure the interstitial fluid glucose level of included patients. It has been verified by the National Institute for Health and Care Excellence based on evidence from randomized controlled trials [46]. Patients were instructed by clinicians to use the device in accordance with the flash glucose monitoring guidelines

suggested by NHS. When using FSL, patients continued to take insulin according to their insulin regimes and type of insulin they use. In addition, patients were arranged to have follow-up consultations every 3 to 6 months, depending on their clinical needs. Pragmatically, the glucose level was primarily measured and recorded once every 15 minutes.

Apart from the FSL data set, CGM data from the REPLACE-BG trial were used for external validation. The REPLACE-BG study is a multicenter randomized trial to evaluate the stand-alone effectiveness of CGM without confirmatory BG measurements in 225 adults with well-controlled T1DM [47]. The trial was chosen for external validation because it represented a patient group that is similar and relevant to this study in 3 ways. First, the REPLACE-BG cohort and our patient cohort both contained patients with T1DM who were using CGM and undergoing similar insulin treatment, which is an important inclusion criterion in this study. Second, the REPLACE-BG trial was conducted in the United States, whereas this study was

conducted in the United Kingdom. The capability of our proposed methods to be applied to patients with different demographics can be tested. Third, as the REPLACE-BG trial included more patients and CGM measurements, it enabled us to validate our methods using a large sample size to demonstrate scalability.

Retrieval and Preprocessing of Clinical Information From Clinical Notes and EPR

The FSL CGM data set did not contain clinical variables that are crucial to this analysis. Thus, clinical notes and EPR were used as sources of clinical information by mapping the participants' NHS identifiers. All available clinical notes between August 5, 2009, and August 5, 2019, were manually reviewed, and the list of medication and diagnosis was extracted for each patient. Then, the list of medication and diagnosis was reviewed by clinicians at the Centre for Diabetes and Endocrinology to categorize them for further analysis (Tables S1 and S2 in [Multimedia Appendix 1](#)). In contrast, the latest laboratory test results, including HbA_{1c} level and estimated glomerular filtration rate, were retrieved from the EPR.

GV Pattern Extraction With DTW

DTW was proposed by Berndt et al [48], and it aims to find patterns in time-series data. The DTW model takes several time-series data as input and outputs the time-series patterns extracted and the type of pattern to which each series belongs. The major strengths of DTW included its ability to handle unevenly spaced time-series data, which is prevalent in CGM data. Several researchers have applied DTW to discover clinical insights such as the prognostic value in CGM data [35], electrocardiograms [49], and genomic signals [50].

A few preprocessing steps were performed to transform the FSL CGM data into inputs for the DTW model. First, if multiple records were found within the same minute in the CGM data, the median value was considered. Second, we divided the CGM data of each patient into overlapping window periods. Any window periods that had <4 measurements per hour on average were discarded to improve model results. Third, hyperparameters of the DTW model, specifically, the duration of each window period and the percentage of overlap between consecutive windows, were tuned. A grid search was performed from a validation set over the 2 hyperparameters to determine the best combination that optimizes a list of cluster validity indexes, namely, Silhouette, Calinski-Harabasz, COP, and modified Davies-Bouldin index. The search space for window duration and overlap percentage were 120, 150, and 180 minutes and 0%, 25%, 50%, and 75%, respectively. The search space for window duration was chosen such that the duration is sufficient to capture the activity profile of rapid-acting insulin.

After determining the aforementioned hyperparameters, the number of patterns to be extracted by the DTW model has to be determined. A DTW model was trained for each of 3 to 8 patterns, and the models were compared. The optimal number of patterns was determined by evaluating the total within-cluster distance against the number of pattern graphs, which is also known as the elbow method. Finally, GV patterns and the type of pattern to which each series belongs were extracted from the

best-performing DTW model. To examine whether our method can be generalized to other CGM data sets on patients with T1DM, we applied the same preprocessing steps and hyperparameters to the REPLACE-BG data set. The number of patterns was determined similarly, and the resulting set of GV patterns was compared with that from FSL data.

Hierarchical Clustering of Patients and Statistical Analysis

Hierarchical clustering algorithm was used to cluster patients with respect to time in patterns, so that no a priori information about the number of clusters would be required [51]. The occurrence of each pattern per patient was tallied and expressed as a percentage of all patterns. Agglomerative hierarchical clustering algorithm with complete linkage was applied on time in patterns, and a dendrogram was plotted. A distance measure specific to percentage data was used for computing the distance matrix for hierarchical clustering instead of the conventional Euclidean distance measure [52]. The number of patient clusters was determined based on the greatest difference in the total within-cluster distance from the dendrogram. Each patient was assigned to one of the clusters for statistical analysis.

In statistical analysis, patient characteristics, including demographics, laboratory test results, diagnoses, and medications, were compared across patient clusters using univariate analysis. Laboratory test results for HbA_{1c} level and estimated glomerular filtration rate were categorized into groups and regarded as categorical variables in 2-tailed statistical tests. ANOVA for continuous variables and chi-square test for categorical or binary variables were performed, and the corresponding *P* values were extracted. Missing values for each variable were omitted from the computation of *P* value. *P* values <.05 were considered as being statistically significant.

Ethics Approval

This study obtained ethics and data governance approval by the Royal Berkshire NHS Foundation Trust under the reference number A2901469.

Results

GV Patterns From DTW Model

A total of 1,590,443 CGM data points across 126 patients was collected in this study. After hyperparameter tuning, it was determined that 150 minutes was the optimal window duration and 50% was the optimal overlap percentage. A comparison of the cluster validity indexes is presented in Figure S1 in [Multimedia Appendix 1](#). This resulted in 149,639 window periods (each 150-minute long) for training the DTW model. By evaluating the graph of the total within-cluster distance against the number of patterns, 6 was found to be the optimal number of patterns. In contrast, GV patterns from the REPLACE-BG data set were extracted with identical configurations, resulting in 931,005 window periods and 5 patterns (Figures S2 and S3 in [Multimedia Appendix 1](#)).

The properties of the 6 GV patterns extracted from the FSL data set are summarized in [Table 2](#). [Figure 2](#) presents several random CGM samples from each pattern group. Results showed that

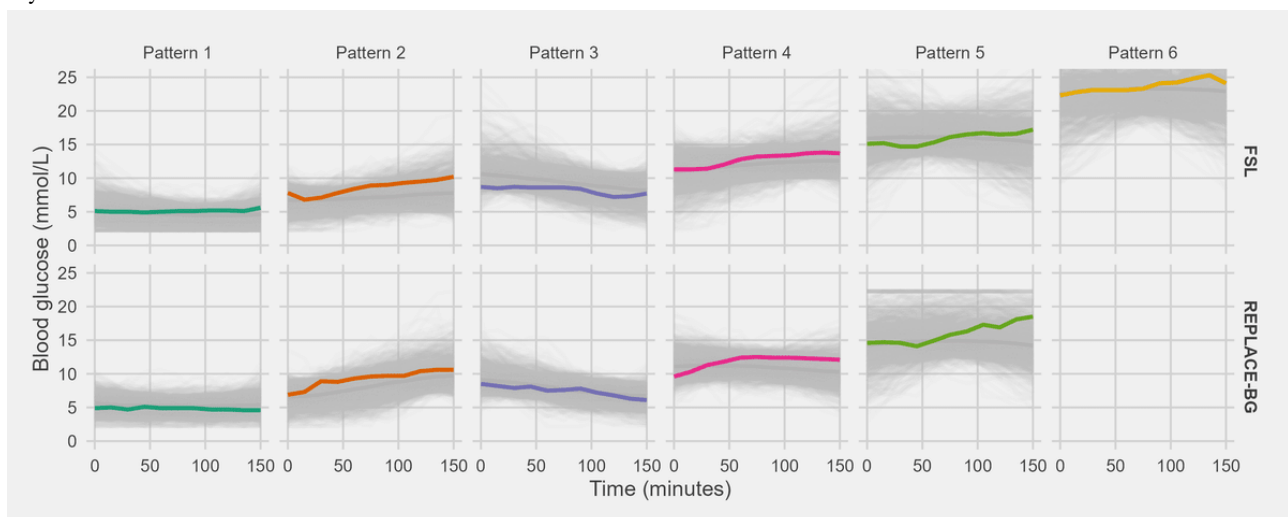
patterns 1 and 2 represent glucose levels at approximately 3 to 6 mmol/L and 6 to 8 mmol/L, respectively, which mostly fall within the target range. A slightly rising trend is also observed in pattern 2. BG trends are also captured in patterns 3 and 4. Pattern 3 represents a decline in BG from marginally hyperglycemic to normal and is the only pattern that depicts an obvious downward trend. In contrast, pattern 4 represents a surge from marginally hyperglycemic to hyperglycemic. Most

of the CGM data belong to patterns 1 to 4, and each of them accounts for approximately 20% of the data. Patterns 5 and 6 both represent less frequent hyperglycemic events at approximately 14 to 19 mmol/L and 19 to 28 mmol/L, respectively. Unlike the other 4 patterns, patterns 5 and 6 had large spread and included different trends that generally falls within their respective glucose levels. In other words, they can include upward, downward, steady, or even parabolic trends.

Table 2. Summary of the 6 glycemic variability (GV) patterns extracted from FreeStyle Libre data set.

GV pattern number	Glucose level	Pattern trends	Occurrence (N=149,639), n (%)
6	Severely hyperglycemic	Steady or rising to peak and declining	8440 (5.64)
5	Hyperglycemic	Steady or concave up or down	22,594 (15.10)
4	From marginally hyperglycemic to hyperglycemic	Rising	28,653 (19.15)
3	From marginally hyperglycemic to normal	Declining	31,185 (20.84)
2	Normal	Steady or slightly rising	30,255 (20.22)
1	Marginally hypoglycemic or normal	Steady or concave up	28,512 (19.05)

Figure 2. Glycemic variability patterns extracted from dynamic time warping model. Each gray line represents a random sample within the specific pattern and data set, and one is highlighted in color. The dark gray line in each panel depicts the median of glycemic variability patterns extracted. FSL: FreeStyle Libre.

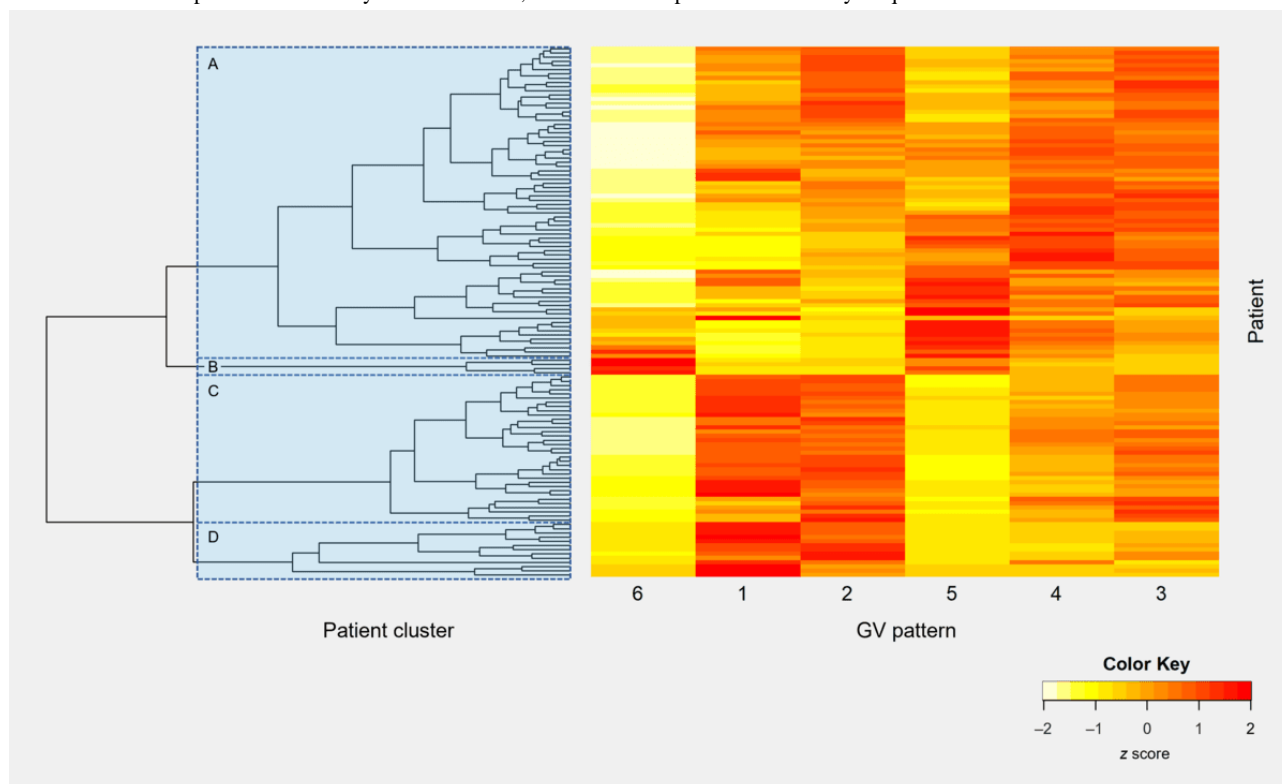


External validation was performed on the REPLACE-BG data set, and results are presented in Figure 2 and Figures S2 and S3 in Multimedia Appendix 1. It is observed that our methods were able to generate a comparable set of GV patterns across the 2 data sets, specifically, patterns 1 to 5. Compared with FSL patterns, the biggest difference in REPLACE-BG patterns is the absence of pattern 6, which indicates severe fluctuations in hyperglycemic events. This is likely owing to the difference in inclusion and exclusion criteria between the 2 data sets. The REPLACE-BG trial cohort deliberately included patients with T1DM who were well controlled and excluded individuals with substantial hypoglycemic events. Therefore, the REPLACE-BG data set is only representative of the well-controlled T1DM group and has limited generalizability to all patients with T1DM. Given that the objective of this study was to generate a comprehensive representation of GV profiles among all patients with T1DM, all further analysis in this study was conducted based on the 6 patterns from FSL data set.

Patient GV Profile Clusters Based on Time in Patterns

Hierarchical clustering was applied on time in GV patterns. Overall, 4 clusters of patients were identified based on the dendrogram (Figure 3). Most patients (74/126, 58.7%) belonged to cluster A. Hyperglycemia fluctuation events occurred more frequently among patients in clusters A and B. Moreover, the time spent in GV patterns 1 and 2 for these patients was relatively low. In particular, the 3.2% (4/126) patients in cluster B spent much more time in GV pattern 6 than in all other clusters. This demonstrates that their glucose level was very poorly controlled and managed. In contrast, the glucose level of patients in clusters C and D are more likely to fall into GV patterns 1 and 2, which roughly resembles the target range. However, patients in cluster C spent relatively more time in GV patterns 3 and 4 when compared with patients in cluster D, which indicates great fluctuation in glucose levels and high likelihood of hyperglycemia events.

Figure 3. Dendrogram in hierarchical clustering and heat map of time in patterns per patient. The left panel depicts the dendrogram in hierarchical clustering. The 4 colored boxes represent 4 different patient clusters based on glycemic variability (GV) patterns. The right panel is a heat map that depicts the underlying distribution of patterns across all patients. Each row represents a patient and each column represents 1 of the 6 extracted GV patterns. Yellow color represents a relatively rare occurrence, and red color represents a relatively frequent occurrence.



Correlation Between GV-Based Clusters and Patient Characteristics

Patient characteristics were compared across the 4 patient clusters and are presented in Table 3. No statistical significance was found across clusters in terms of demographical variables, except for age ($P=.02$). Specifically, patients in cluster B were observed to be younger and had shorter duration of diabetes than those in the other 3 clusters ($P=.04$). Moreover, the patient clusters were significantly different in various glycemic metrics, including HbA_{1c} level category ($P<.001$), COV ($P=.003$), and

TIR ($P<.002$). Patients in cluster D were associated with high odds of meeting the HbA_{1c} level and TIR recommended targets. Although more than half of patients in cluster C (23/35, 66%) met the recommended target for HbA_{1c} level, they had one of the greatest COV among all 4 clusters, and only 11% (4/35) of them met the recommended target for COV. Patients in clusters A and B were associated with significantly increased likelihood of poorly controlled diabetes. Most patients in cluster A and all patients in cluster B failed to fulfill HbA_{1c} level (7/74, 10%) and TIR targets, indicating further management needs in terms of type or dosage of insulin intake.

Table 3. Patient characteristics across the 4 patient clusters (N=126).

Characteristics	Cluster A (n=74)	Cluster B (n=4)	Cluster C (n=35)	Cluster D (n=13)	P value ^a
Age (years), mean (SD)	40.3 (14.3)	22.8 (4.27)	41.8 (12.9)	33.8 (10.7)	.02
Sex (female), n (%)	34 (46)	1 (25)	18 (51)	9 (69)	.33
Index of Multiple Deprivation decile [53], mean (SD)	7.96 (2.31)	8.25 (2.06)	7.34 (2.44)	7.38 (2.66)	.56
BMI (kg/m ²), mean (SD)	27.3 (4.64)	24.1 (3.07)	26.9 (5.69)	23.4 (4.39)	.22
Duration of diabetes (years), mean (SD)	22.2 (12.4)	8.5 (4.36)	24.8 (15.7)	14.5 (14.7)	.04
Number of days since CGM ^b use, mean (SD)	218 (243)	203 (56.1)	167 (203)	215 (276)	.76
eGFR^c stage, n (%)					.85
5	0 (0)	0 (0)	0 (0)	0 (0)	
4	1 (1)	0 (0)	0 (0)	0 (0)	
3b	2 (3)	0 (0)	0 (0)	0 (0)	
3a	4 (5)	0 (0)	1 (3)	0 (0)	
2	28 (38)	0 (0)	16 (46)	6 (46)	
1	38 (51)	4 (100)	18 (51)	7 (54)	
HbA_{1c}^d level (mmol/mol), n (%)					<.001
≤42	1 (1)	0 (0)	1 (3)	7 (54)	
43-48	4 (5)	0 (0)	6 (17)	3 (23)	
48-59	14 (19)	0 (0)	19 (54)	2 (15)	
59-85	46 (62)	1 (25)	9 (26)	1 (8)	
≥86	8 (11)	3 (75)	0 (0)	0 (0)	
Glucose level, mean (SD)	10.8 (1.57)	19.3 (1.46)	8.22 (0.614)	6.48 (1.1)	<.001
COV ^e of glucose level, mean (SD)	0.428 (0.066)	0.354 (0.031)	0.429 (0.061)	0.37 (0.063)	.003
TIR^f (mmol/L), mean % (SD)					
≤3	1.9 (2.2)	0.3 (0.2)	3.2 (2.8)	4.6 (5)	.002
3-3.9	3.5 (2)	0.6 (0.2)	6.7 (2.7)	11 (7.7)	<.001
3.9-10	43.3 (11.3)	10.6 (2.9)	62.4 (6.4)	74.7 (12.9)	<.001
10-13.9	27.2 (6.0)	13.3 (3.6)	20.5 (4)	7.5 (4.9)	<.001
≥13.9	24.1 (11.6)	75.2 (5.8)	7.3 (3.4)	2.3 (5.8)	<.001
Time in patterns, mean % (SD)					
1	13.6 (7.2)	2.1 (0.6)	27.6 (8.8)	51.9 (17.9)	<.001
2	17 (6.1)	3.6 (1.1)	27.6 (5)	32.5 (10.6)	<.001
3	22.4 (4.7)	6.8 (1.7)	23.5 (5.2)	10.2 (6.6)	<.001
4	23.1 (5.7)	10.1 (2.8)	15.6 (4.2)	5 (7.3)	<.001
5	19.3 (7.8)	26.2 (5.8)	5.5 (3)	0.4 (0.7)	<.001
6	4.7 (6)	51.2 (11)	0.3 (0.4)	0 (0)	<.001
Fulfillment of recommended targets [28], n (%)					
TIR between 3.9 and 10 mmol/L >70% of the time	7 (10)	0 (0)	23 (66)	11 (85)	<.001
COV of glucose level <0.36	8 (11)	2 (50)	4 (11)	7 (54)	<.001
HbA _{1c} level <58 mmol/mol	18 (24)	0 (0)	26 (74)	12 (92)	<.001
Comorbidities, n (%)	55 (74)	3 (75)	22 (63)	5 (39)	.10

Characteristics	Cluster A (n=74)	Cluster B (n=4)	Cluster C (n=35)	Cluster D (n=13)	<i>P</i> value ^a
Diabetic complications, n (%)	47 (64)	3 (75)	19 (54)	4 (31)	.18
Medication, n (%)					
Insulin					
Injection	66 (89)	4 (100)	33 (94)	12 (92)	.75
Pump	11 (15)	0 (0)	4 (11)	2 (15)	.78
Blood pressure	19 (26)	1 (25)	9 (26)	1 (8)	.55
Cholesterol	19 (26)	0 (0)	8 (23)	1 (8)	.33
Thyroid	10 (14)	0 (0)	2 (6)	2 (15)	.50
Antiplatelet	4 (5)	0 (0)	2 (6)	1 (8)	.95
Psychology	7 (10)	0 (0)	0 (0)	0 (0)	.15

^a*P* values <.05 are italicized; missing values were omitted only during the calculation of *P* values.

^bCGM: continuous glucose monitoring.

^ceGFR: estimated glomerular filtration rate.

^dHbA_{1c}: glycated hemoglobin.

^eCOV: coefficient of variation.

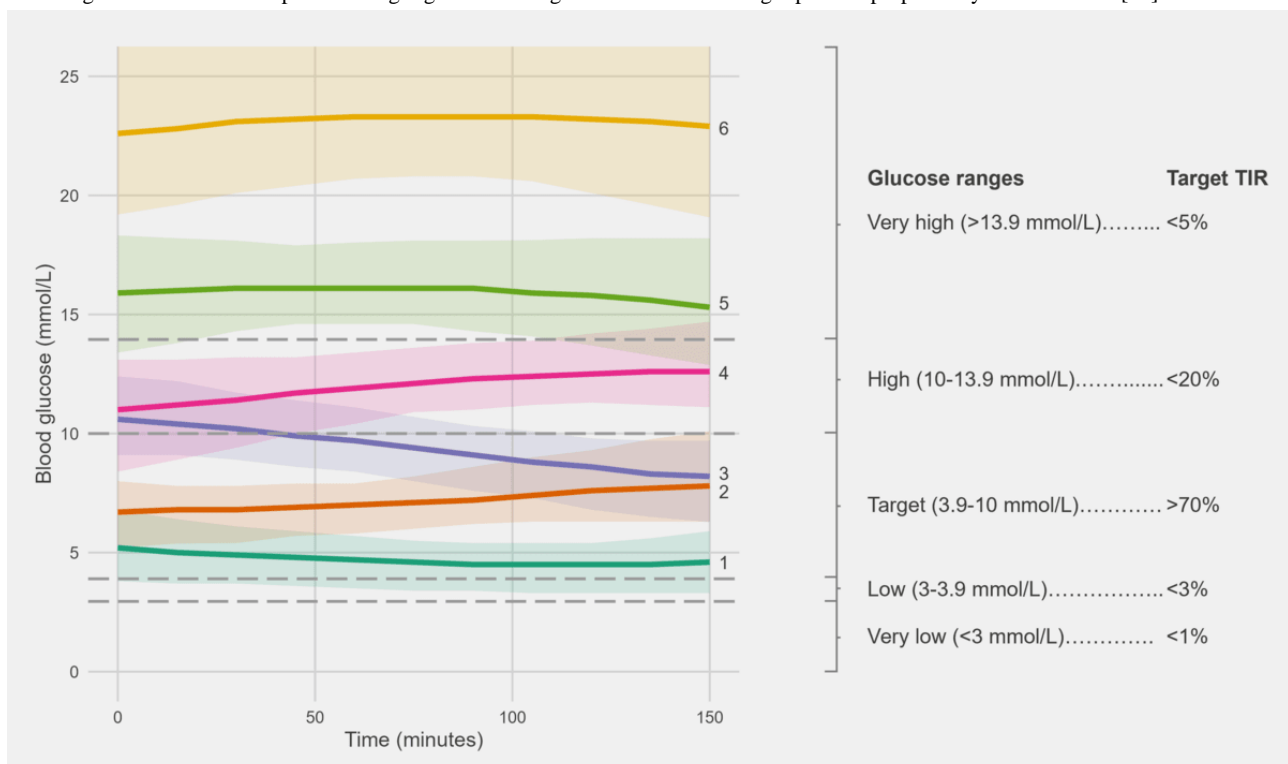
^fTIR: time in range.

Resemblance Between TIRs and Time in GV Patterns

It is possible to translate some of the TIR targets to targets of GV patterns owing to their conceptual similarity, and it is observed that some of the extracted GV patterns resemble the TIR glucose cutoff points as recommended by Battelino et al [28] (Figure 4). This can potentially serve as reference to better understand the clinical impacts for each pattern. GV patterns 5 and 6 both belong to the very high glucose range. Thus, a

recommended target TIR of <5% within the very high glucose range can be approximately translated to having <5% occurrence for patterns 5 and 6. Pattern 4 generally represents high glucose level, with cutoffs at approximately 10 and 13.9 mmol/L. However, none of the patterns exclusively covers the very low glucose range (<3.9 mmol/L). This is because such readings were very rare in the data set, such that they were inherently grouped into GV pattern 1 by the DTW model.

Figure 4. Comparison of recommended time in range (TIR) targets and extracted glycemic variability patterns. Each color in the left panel represents a glycemic variability pattern. The lower and upper bound of each shaded region represent the 20th and 80th percentile of glucose trend for that pattern. The median glucose trend of each pattern is highlighted. The target TIR shown in the right panel is proposed by Battelino et al [28].



As our extracted GV patterns take fluctuation in BG into account in addition to its magnitude, our method is able to provide additional context for a person's BG profile. The prevalence of GV patterns 4 and 5 would indicate a fluctuation between high and very high glucose ranges, whereas that of GV patterns 3 and 4 indicates a fluctuation between target to high glucose level. This piece of information cannot be deduced from TIR. It should be noted that taking fluctuation into account also implies that direct translation from TIR targets to certain patterns is unavailable, as they span across different glucose ranges. For instance, the target glucose level ranges between 3.9 and 10 mmol/L comprises patterns 1, 2, and 3.

GV Patterns Over Time

In this study, we sought to draw insights about patients with different time in GV patterns by using hierarchical clustering. A total of 4 clusters was found, each with very distinguishing glycemetic fluctuation features and thus management needs. An example of daily glucose trends from each cluster is presented in [Figure 5](#). Diabetes in patients in cluster D was well controlled, and there is no need to alter their insulin regime. Although the glucose level of patients in cluster C usually falls within target range, it has great variability, which could indicate the need for changing their insulin regimes to reduce fluctuation and hyperglycemia events. In contrast, patients in clusters A and B had very poorly controlled diabetes, and a significant increase in fluctuation severity is observed, which suggests the need for change in glucose management. Patients in cluster A show sharp increases and decreases across target and hyperglycemia ranges, whereas those in cluster B primarily fluctuate at hyperglycemia level. A possible explanation for this is that patients in cluster B tend to be young and had short duration of diabetes. Therefore, the optimal way to manage their glucose levels is less apparent and would still require some time to be determined in follow-up consultations. Apart from existing metrics such as HbA_{1c} level

and TIR, we believe that studying patient clusters can be beneficial as a complementary metric during consultations, which could improve patient care and, ultimately, clinical outcomes.

To better understand the properties of each GV pattern, we further evaluated the relationship between GV patterns and time of day. The occurrence of patterns across time of day according to cluster is presented in [Figure 6](#). It is observed that GV pattern 1, which represents steady glucose level around marginal hypoglycemia to normal, most frequently occurs at midnight between 2 AM and 6 AM. This is likely owing to the absence of food intake during the period. In contrast, GV patterns 2 and 4, which are indicators of a surge in glucose level, are more likely to occur at typical meal hours around 9 AM, 1 PM, and 7 PM for patients in clusters C and D. Similarly, GV patterns 5 and 6 occur the most within that period for patients in cluster B whose glucose level are very poorly controlled. These observations are generally consistent with existing literature about the daily fluctuation in glucose levels [29].

Apart from analyzing GV patterns over time of day, we further investigated whether the duration of CGM use is associated with patients' GV profile and characteristics. On the basis of the distribution of CGM use duration in our data set, the cohort is divided into 3 approximately equal-sized groups to facilitate comparison: <68 days (46/126, 36.5%), 68 to 180 days (40/126, 31.7%), and >180 days (40/126, 31.7%). Our findings revealed that although no statistical significance was found between the duration of CGM use and patient demographics or fulfillment of recommended glycemetic targets (all $P > .05$; [Table 4](#)), long duration is associated with specific glycemetic metrics, including high mean glucose ($P = .03$), TIR ≥ 13.9 mmol/L ($P = .04$), and time in pattern 6 ($P = .04$). This may indicate increased likelihood of poorly controlled or managed patients who have been using CGM for an extended period.

Figure 5. The 1-day glucose trend of patients sampled from each cluster. The shaded region represents the target glucose range, and the 6 glycemic variability (GV) patterns over time are highlighted in 6 colors.

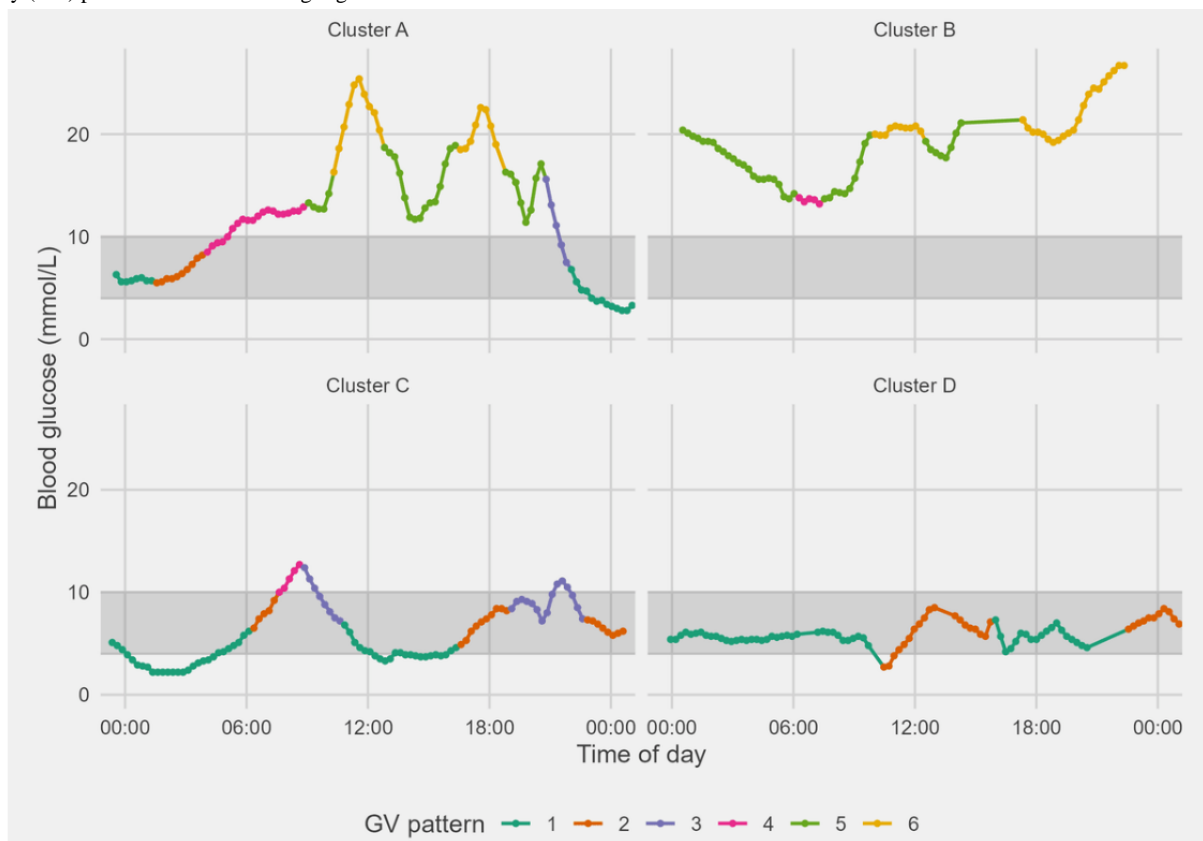


Figure 6. Hourly distribution of glycemic variability (GV) patterns across a day for each patient cluster.

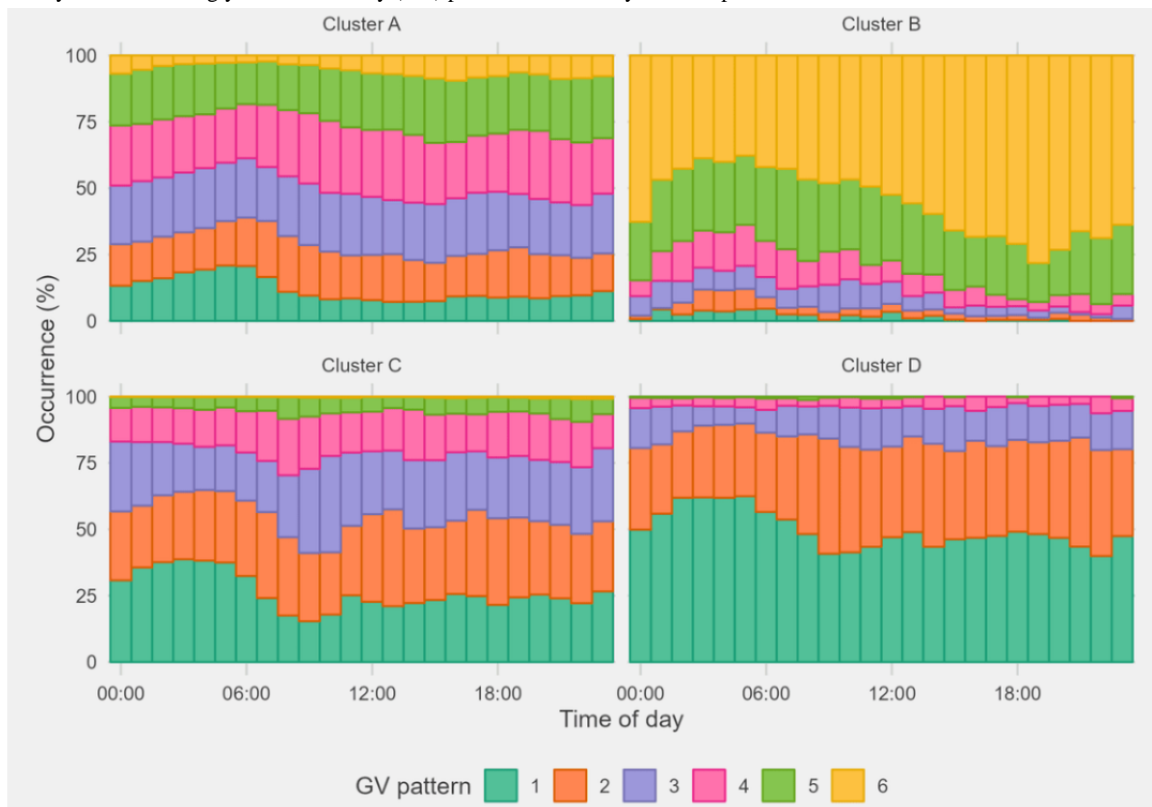


Table 4. Patient characteristics across different duration of diabetes (N=126).

Characteristics	<68 days (n=46)	68-180 days (n=40)	>180 days (n=40)	<i>P</i> value ^a
Age (years), mean (SD)	38.7 (12.5)	40.4 (15.2)	39.4 (14)	.87
Sex (female), n (%)	26 (57)	16 (40)	20 (50)	.31
Index of Multiple Deprivation decile [53], mean (SD)	7.83 (2.01)	8.1 (2.34)	7.28 (2.73)	.28
BMI (kg/m ²), mean (SD)	27.8 (6.2)	26.8 (3.9)	25.7 (4)	.22
Glucose level, mean (SD)	9.08 (1.57)	10.2 (2.5)	10.5 (3.5)	.03
COV ^b of glucose level, mean (SD)	0.434 (0.074)	0.416 (0.06)	0.408 (0.061)	.17
TIR^c (mmol/L), mean % (SD)				
≤3	3 (3.5)	1.7 (1.5)	2.6 (3)	.11
3-3.9	5.9 (3.6)	4.3 (2.8)	5 (5.3)	.19
3.9-10	55.3 (13.5)	49.1 (17.3)	47.3 (18.9)	.07
10-13.9	21.7 (7.2)	24.3 (8.5)	22.8 (9)	.36
≥13.9	14.1 (10.3)	20.6 (15.8)	22.4 (21.1)	.04
Time in patterns, mean % (SD)				
1	24.7 (14.5)	17.4 (13.5)	20.5 (17.5)	.09
2	23.4 (8.8)	20.2 (8.8)	19.4 (9.7)	.10
3	21.2 (7)	21.7 (6.3)	19.9 (7.1)	.46
4	17.8 (7.8)	20.1 (8.2)	18.3 (8.3)	.40
5	11.1 (9.2)	15.6 (9.4)	14.8 (11.1)	.08
6	1.7 (2.7)	4.9 (8.7)	7.1 (14.9)	.04
Fulfillment of recommended targets [28], n (%)				
TIR between 3.9 and 10 mmol/L >70% of the time	17 (37)	11 (28)	13 (33)	.65
COV of glucose level <0.36	7 (15)	8 (20)	6 (15)	.79
HbA _{1c} ^d level <58 mmol/mol	19 (41)	17 (43)	20 (50)	.61

^a*P* values <.05 are italicized; missing values were omitted only during the calculation of *P* values.

^bCOV: coefficient of variation.

^cTIR: time in range.

^dHbA_{1c}: glycated hemoglobin.

Discussion

Principal Findings

As an important application of smart and connected health, CGM has been gaining popularity rapidly ever since its inception and is becoming a vital tool to improve glucose management in patients with T1DM. With the increasing use of CGM for managing patients with T1DM, metrics such as TIR are recommended to depict GV, but a significant part of information available in CGM data is often omitted. In this study, we proposed a machine learning framework for extracting GV patterns from CGM data that harnesses the strengths of machine learning in terms of the capability of analyzing large amounts of data. By applying DTW on CGM data, we showed that it is possible to extract recurring patterns in CGM that inherit the clinical concepts of TIR, a recognized CGM-derived metric. Specifically, 6 distinctive patterns were found, and we showed

that time in patterns can be used to comprehensively represent patients' GV profile and to complement TIR owing to their conceptual resemblance. We further drew insights from GV patterns by identifying the types of patients with T1DM based on time in patterns and addressing the relationship between GV patterns and time of day. Our method captured information beyond absolute glucose value and revealed the details of glucose variability and dynamics. We demonstrated that time in patterns is an accessible, more comprehensive representation of a patient's GV and could provide additional insights such as types of patients with T1DM and time of day.

Our proposed methods successfully captured GV patterns that inherently incorporate the idea of clinically meaningful concepts such as mean glucose level, GV, and TIR. Time in patterns derived from our methods contains much rich information, as existing methods such as TIR disregard the sequence in which the glucose measurements were made. Finally, an advantage of

our time-in-patterns method over other proposed machine learning-based metrics is its scalability and understandability, which is largely owing to the ability to visualize our extracted patterns from blood monitoring data. As mentioned in section *Quantifying GV*, clinical understandability is a major issue that hindered machine learning-based GV extraction methods from being a widely accepted glycemic metric. For example, it is generally more meaningful to portray GV using time in patterns, such as 36% time spent in GV pattern 3 (rising from marginally hyperglycemic to normal) and pattern 4 (declining from marginally hyperglycemic to hyperglycemic), than a single SD value such as 0.36. We also validated the blood fluctuation patterns 1 to 5 using US-based CGM data from the REPLACE-BG trial of 225 adults with well-controlled T1DM. This shows that our method has the generalizability to cover different patient cohorts from various demographics.

Limitations

This study had a few limitations. First, the duration of CGM use varied from 1 month to 3 years across patients in this study. Although no significant association was found between days since the use of CGM and patient cluster ($P=.76$), certain effects may not be accounted for in this study, such as seasonal effects on glucose levels [54]. Second, the adoption of CGM at the moment is still limited to the well-developed areas of the world where there are information and communication technology infrastructure with high level of digital readiness for connected health and sufficient funding for patients with T1DM to use wearable CGM devices. This is also reflected in our data that the patients included in this study were predominantly living in less deprived areas. For example, 75.4% (95/126) of the

patients in our study were living in less deprived areas according to the Index of Multiple Deprivation (IMD) decile ($IMD \geq 7$), and 34.9% (44/126) of them were living in the least deprived area ($IMD=10$). Only 19.8% (25/126) of the patients in our study were living in more deprived areas ($IMD \leq 5$). The average IMD decile in different patient clusters can be found in Table 3. Therefore, the generalizability of our results to other demographics such as patients living in rural areas is limited. It should also be noted that apart from infrastructure and deprivation, there are other factors affecting the adoption of CGM such as device accuracy [55], user perception, device obtrusiveness [56], and interpersonal influence [57]. Third, as only the latest list of medication and laboratory test results was collected from each patient, any change in medication or management throughout the study period was not accounted for. A patient who spent a lot of time in hyperglycemia may remain in the target glucose range steadily after a change in their insulin regime. In this case, the resulting time in patterns would be averaged across the 2 states and fail to represent the patient's latest situation.

Future Studies

Future studies could focus on investigating the clinical relationship between GV patterns and DM medications through prospective studies and randomized control trials. By having a more comprehensive representation of GV profile, we can better categorize patients, which in turn would enable us to understand more about their unique condition and needs. We believe that this framework can ultimately serve as a step toward the development of personalized therapeutic pathways for patients with DM in the environment of connected health.

Data Availability

The anonymized data set of this study was collected from the Centre for Diabetes and Endocrinology at the Royal Berkshire National Health Service Foundation Trust, the United Kingdom. The data set may be available subject to ethics and information governance approval from Royal Berkshire National Health Service Foundation Trust. The validation data set is from the REPLACE-BG randomized trial conducted in the United States among adults with well-controlled type 1 diabetes [47].

Authors' Contributions

WL and TA developed the idea of the study. NBC performed the experiments, developed the computational programs, and played a major part in drafting the initial manuscript. WL and TA supervised the project. EB supported the data management. RMM advised on the scientific content of the project. All authors participated in producing the final manuscript draft and approved the final submitted version.

Conflicts of Interest

None declared.

Multimedia Appendix 1

Supplemental tables and figures.

[PDF File (Adobe PDF File), 473 KB - ai_v2i1e45450_app1.pdf]

References

1. IDF Diabetes Atlas 10th edition. International Diabetes Federation. 2021. URL: <https://diabetesatlas.org/atlas/tenth-edition/> [accessed 2022-09-26]

2. Global health estimates: Leading causes of death. Cause-specific mortality, 2000–2019. World Health Organization. 2020. URL: <https://www.who.int/data/gho/data/themes/mortality-and-global-health-estimates/gho-leading-causes-of-death> [accessed 2022-09-26]
3. Juvenile Diabetes Research Foundation Continuous Glucose Monitoring Study Group, Tamborlane WV, Beck RW, Bode BW, Buckingham B, Chase HP, et al. Continuous glucose monitoring and intensive treatment of type 1 diabetes. *N Engl J Med* 2008 Oct 02;359(14):1464-1476. [doi: [10.1056/NEJMoa0805017](https://doi.org/10.1056/NEJMoa0805017)] [Medline: [18779236](https://pubmed.ncbi.nlm.nih.gov/18779236/)]
4. Reddy N, Verma N, Dungan K. Monitoring technologies- continuous glucose monitoring, mobile technology, biomarkers of glycemic control. *Endotext*. 2020 Aug 16. URL: <https://www.ncbi.nlm.nih.gov/books/NBK279046/> [accessed 2022-09-26]
5. Foster NC, Beck RW, Miller KM, Clements MA, Rickels MR, DiMeglio LA, et al. State of type 1 diabetes management and outcomes from the T1D exchange in 2016-2018. *Diabetes Technol Ther* 2019 Feb;21(2):66-72 [FREE Full text] [doi: [10.1089/dia.2018.0384](https://doi.org/10.1089/dia.2018.0384)] [Medline: [30657336](https://pubmed.ncbi.nlm.nih.gov/30657336/)]
6. Janapala RN, Jayaraj JS, Fathima N, Kashif T, Usman N, Dasari A, et al. Continuous glucose monitoring versus self-monitoring of blood glucose in type 2 diabetes mellitus: a systematic review with meta-analysis. *Cureus* 2019 Sep 12;11(9):e5634 [FREE Full text] [doi: [10.7759/cureus.5634](https://doi.org/10.7759/cureus.5634)] [Medline: [31700737](https://pubmed.ncbi.nlm.nih.gov/31700737/)]
7. Huisman TH, Martis EA, Dozy A. Chromatography of hemoglobin types on carboxymethylcellulose. *J Lab Clin Med* 1958 Aug;52(2):312-327. [Medline: [13564011](https://pubmed.ncbi.nlm.nih.gov/13564011/)]
8. Bookchin RM, Gallop PM. Structure of hemoglobin A1c: nature of the N-terminal beta chain blocking group. *Biochem Biophys Res Commun* 1968 Jul 11;32(1):86-93. [doi: [10.1016/0006-291x\(68\)90430-0](https://doi.org/10.1016/0006-291x(68)90430-0)] [Medline: [4874776](https://pubmed.ncbi.nlm.nih.gov/4874776/)]
9. Nathan DM, Turgeon H, Regan S. Relationship between glycosylated haemoglobin levels and mean glucose levels over time. *Diabetologia* 2007 Nov;50(11):2239-2244 [FREE Full text] [doi: [10.1007/s00125-007-0803-0](https://doi.org/10.1007/s00125-007-0803-0)] [Medline: [17851648](https://pubmed.ncbi.nlm.nih.gov/17851648/)]
10. Dhataria K, Levy N, Kilvert A, Watson B, Cousins D, Flanagan D, Joint British Diabetes Societies. NHS diabetes guideline for the perioperative management of the adult patient with diabetes. *Diabet Med* 2012 Apr;29(4):420-433. [doi: [10.1111/j.1464-5491.2012.03582.x](https://doi.org/10.1111/j.1464-5491.2012.03582.x)] [Medline: [22288687](https://pubmed.ncbi.nlm.nih.gov/22288687/)]
11. American Diabetes Association. Standards of medical care in diabetes-2019 abridged for primary care providers. *Clin Diabetes* 2019 Jan;37(1):11-34 [FREE Full text] [doi: [10.2337/cd18-0105](https://doi.org/10.2337/cd18-0105)] [Medline: [30705493](https://pubmed.ncbi.nlm.nih.gov/30705493/)]
12. Wilmot EG, Lumb A, Hammond P, Murphy HR, Scott E, Gibb FW, et al. Time in range: a best practice guide for UK diabetes healthcare professionals in the context of the COVID-19 global pandemic. *Diabet Med* 2021 Jan;38(1):e14433 [FREE Full text] [doi: [10.1111/dme.14433](https://doi.org/10.1111/dme.14433)] [Medline: [33073388](https://pubmed.ncbi.nlm.nih.gov/33073388/)]
13. Suh S, Kim JH. Glycemic variability: how do we measure it and why is it important? *Diabetes Metab J* 2015 Aug;39(4):273-282 [FREE Full text] [doi: [10.4093/dmj.2015.39.4.273](https://doi.org/10.4093/dmj.2015.39.4.273)] [Medline: [26301188](https://pubmed.ncbi.nlm.nih.gov/26301188/)]
14. Beck RW, Connor CG, Mullen DM, Wesley DM, Bergenstal RM. The fallacy of average: how using HbA1c alone to assess glycemic control can be misleading. *Diabetes Care* 2017 Aug;40(8):994-999 [FREE Full text] [doi: [10.2337/dc17-0636](https://doi.org/10.2337/dc17-0636)] [Medline: [28733374](https://pubmed.ncbi.nlm.nih.gov/28733374/)]
15. Cardoso CR, Leite NC, Moram CB, Salles GF. Long-term visit-to-visit glycemic variability as predictor of micro- and macrovascular complications in patients with type 2 diabetes: the Rio de Janeiro Type 2 Diabetes Cohort Study. *Cardiovasc Diabetol* 2018 Feb 24;17(1):33 [FREE Full text] [doi: [10.1186/s12933-018-0677-0](https://doi.org/10.1186/s12933-018-0677-0)] [Medline: [29477146](https://pubmed.ncbi.nlm.nih.gov/29477146/)]
16. Hirsch IB, Brownlee M. Should minimal blood glucose variability become the gold standard of glycemic control? *J Diabetes Complications* 2005 May;19(3):178-181. [doi: [10.1016/j.jdiacomp.2004.10.001](https://doi.org/10.1016/j.jdiacomp.2004.10.001)] [Medline: [15866065](https://pubmed.ncbi.nlm.nih.gov/15866065/)]
17. Bellaver P, Schaeffer AF, Dullius DP, Viana MV, Leitão CB, Rech TH. Association of multiple glycemic parameters at intensive care unit admission with mortality and clinical outcomes in critically ill patients. *Sci Rep* 2019 Dec 06;9(1):18498 [FREE Full text] [doi: [10.1038/s41598-019-55080-3](https://doi.org/10.1038/s41598-019-55080-3)] [Medline: [31811218](https://pubmed.ncbi.nlm.nih.gov/31811218/)]
18. Todi S, Bhattacharya M. Glycemic variability and outcome in critically ill. *Indian J Crit Care Med* 2014 May;18(5):285-290 [FREE Full text] [doi: [10.4103/0972-5229.132484](https://doi.org/10.4103/0972-5229.132484)] [Medline: [24914256](https://pubmed.ncbi.nlm.nih.gov/24914256/)]
19. Krinsley JS. Glycemic variability: a strong independent predictor of mortality in critically ill patients. *Crit Care Med* 2008 Nov;36(11):3008-3013. [doi: [10.1097/CCM.0b013e31818b38d2](https://doi.org/10.1097/CCM.0b013e31818b38d2)] [Medline: [18824908](https://pubmed.ncbi.nlm.nih.gov/18824908/)]
20. Oh TK, Heo E, Song IA, Jeong WJ, Han M, Bang JS. Increased glucose variability during long-term therapeutic hypothermia as a predictor of poor neurological outcomes and mortality: a retrospective study. *Ther Hypothermia Temp Manag* 2020 Jun;10(2):106-113. [doi: [10.1089/ther.2019.0004](https://doi.org/10.1089/ther.2019.0004)] [Medline: [31161969](https://pubmed.ncbi.nlm.nih.gov/31161969/)]
21. Gorst C, Kwok CS, Aslam S, Buchan I, Kontopantelis E, Myint PK, et al. Long-term glycemic variability and risk of adverse outcomes: a systematic review and meta-analysis. *Diabetes Care* 2015 Dec;38(12):2354-2369. [doi: [10.2337/dc15-1188](https://doi.org/10.2337/dc15-1188)] [Medline: [26604281](https://pubmed.ncbi.nlm.nih.gov/26604281/)]
22. Service FJ. Glucose variability. *Diabetes* 2013 May;62(5):1398-1404 [FREE Full text] [doi: [10.2337/db12-1396](https://doi.org/10.2337/db12-1396)] [Medline: [23613565](https://pubmed.ncbi.nlm.nih.gov/23613565/)]
23. Siegelar SE, Holleman F, Hoekstra JB, DeVries JH. Glucose variability; does it matter? *Endocr Rev* 2010 Apr;31(2):171-182. [doi: [10.1210/er.2009-0021](https://doi.org/10.1210/er.2009-0021)] [Medline: [19966012](https://pubmed.ncbi.nlm.nih.gov/19966012/)]
24. Omar AS, Salama A, Allam M, Elgohary Y, Mohammed S, Tuli AK, et al. Association of time in blood glucose range with outcomes following cardiac surgery. *BMC Anesthesiol* 2015 Jan 26;15(1):14 [FREE Full text] [doi: [10.1186/1471-2253-15-14](https://doi.org/10.1186/1471-2253-15-14)] [Medline: [25670921](https://pubmed.ncbi.nlm.nih.gov/25670921/)]

25. Beck RW, Bergenstal RM, Riddlesworth TD, Kollman C, Li Z, Brown AS, et al. Validation of time in range as an outcome measure for diabetes clinical trials. *Diabetes Care* 2019 Mar;42(3):400-405 [FREE Full text] [doi: [10.2337/dc18-1444](https://doi.org/10.2337/dc18-1444)] [Medline: [30352896](https://pubmed.ncbi.nlm.nih.gov/30352896/)]
26. Lu J, Ma X, Zhou J, Zhang L, Mo Y, Ying L, et al. Association of time in range, as assessed by continuous glucose monitoring, with diabetic retinopathy in type 2 diabetes. *Diabetes Care* 2018 Nov;41(11):2370-2376. [doi: [10.2337/dc18-1131](https://doi.org/10.2337/dc18-1131)] [Medline: [30201847](https://pubmed.ncbi.nlm.nih.gov/30201847/)]
27. Beyond A1C Writing Group. Need for regulatory change to incorporate beyond A1C glycemic metrics. *Diabetes Care* 2018 Jun;41(6):e92-e94. [doi: [10.2337/dci18-0010](https://doi.org/10.2337/dci18-0010)] [Medline: [29784704](https://pubmed.ncbi.nlm.nih.gov/29784704/)]
28. Battelino T, Danne T, Bergenstal RM, Amiel SA, Beck R, Biester T, et al. Clinical targets for continuous glucose monitoring data interpretation: recommendations from the international consensus on time in range. *Diabetes Care* 2019 Aug;42(8):1593-1603 [FREE Full text] [doi: [10.2337/dci19-0028](https://doi.org/10.2337/dci19-0028)] [Medline: [31177185](https://pubmed.ncbi.nlm.nih.gov/31177185/)]
29. Advani A. Positioning time in range in diabetes management. *Diabetologia* 2020 Feb;63(2):242-252. [doi: [10.1007/s00125-019-05027-0](https://doi.org/10.1007/s00125-019-05027-0)] [Medline: [31701199](https://pubmed.ncbi.nlm.nih.gov/31701199/)]
30. Murphy HR. Continuous glucose monitoring targets in type 1 diabetes pregnancy: every 5% time in range matters. *Diabetologia* 2019 Jul;62(7):1123-1128 [FREE Full text] [doi: [10.1007/s00125-019-4904-3](https://doi.org/10.1007/s00125-019-4904-3)] [Medline: [31161344](https://pubmed.ncbi.nlm.nih.gov/31161344/)]
31. Struble N. *Measuring Glycemic Variability and Predicting Blood Glucose Levels: Using Machine Learning Regression Models*. London, UK: LAP LAMBERT Academic Publishing; Dec 2013.
32. Marling CR, Struble NW, Bunescu RC, Shubrook JH, Schwartz FL. A consensus perceived glycemic variability metric. *J Diabetes Sci Technol* 2013 Jul 01;7(4):871-879 [FREE Full text] [doi: [10.1177/193229681300700409](https://doi.org/10.1177/193229681300700409)] [Medline: [23911168](https://pubmed.ncbi.nlm.nih.gov/23911168/)]
33. Eljil KS, Qadah GZ, Pasquier M. Predicting hypoglycemia in diabetic patients using time-sensitive artificial neural networks. *Int J Healthc Inf Syst Inform* 2016 Oct;11(4):70-88 [FREE Full text] [doi: [10.4018/ijhisi.2016100104](https://doi.org/10.4018/ijhisi.2016100104)]
34. Mani S, Chen Y, Elasy T, Clayton W, Denny J. Type 2 diabetes risk forecasting from EMR data using machine learning. *AMIA Annu Symp Proc* 2012;2012:606-615 [FREE Full text] [Medline: [23304333](https://pubmed.ncbi.nlm.nih.gov/23304333/)]
35. Hall H, Perelman D, Breschi A, Limcaoco P, Kellogg R, McLaughlin T, et al. Glucotypes reveal new patterns of glucose dysregulation. *PLoS Biol* 2018 Jul 24;16(7):e2005143 [FREE Full text] [doi: [10.1371/journal.pbio.2005143](https://doi.org/10.1371/journal.pbio.2005143)] [Medline: [30040822](https://pubmed.ncbi.nlm.nih.gov/30040822/)]
36. Umpierrez GE, P Kovatchev B. Glycemic variability: how to measure and its clinical implication for type 2 diabetes. *Am J Med Sci* 2018 Dec;356(6):518-527 [FREE Full text] [doi: [10.1016/j.amjms.2018.09.010](https://doi.org/10.1016/j.amjms.2018.09.010)] [Medline: [30447705](https://pubmed.ncbi.nlm.nih.gov/30447705/)]
37. Rodbard D. Clinical interpretation of indices of quality of glycemic control and glycemic variability. *Postgrad Med* 2011 Jul;123(4):107-118. [doi: [10.3810/pgm.2011.07.2310](https://doi.org/10.3810/pgm.2011.07.2310)] [Medline: [21680995](https://pubmed.ncbi.nlm.nih.gov/21680995/)]
38. Rama Chandran S, Tay WL, Lye WK, Lim LL, Ratnasingam J, Tan AT, et al. Beyond HbA1c: comparing glycemic variability and glycemic indices in predicting hypoglycemia in type 1 and type 2 diabetes. *Diabetes Technol Ther* 2018 May;20(5):353-362. [doi: [10.1089/dia.2017.0388](https://doi.org/10.1089/dia.2017.0388)] [Medline: [29688755](https://pubmed.ncbi.nlm.nih.gov/29688755/)]
39. McDonnell CM, Donath SM, Vidmar SI, Werther GA, Cameron FJ. A novel approach to continuous glucose analysis utilizing glycemic variation. *Diabetes Technol Ther* 2005 Apr;7(2):253-263. [doi: [10.1089/dia.2005.7.253](https://doi.org/10.1089/dia.2005.7.253)] [Medline: [15857227](https://pubmed.ncbi.nlm.nih.gov/15857227/)]
40. Service FJ, Molnar GD, Rosevear JW, Ackerman E, Gatewood LC, Taylor WF. Mean amplitude of glycemic excursions, a measure of diabetic instability. *Diabetes* 1970 Sep;19(9):644-655. [doi: [10.2337/diab.19.9.644](https://doi.org/10.2337/diab.19.9.644)] [Medline: [5469118](https://pubmed.ncbi.nlm.nih.gov/5469118/)]
41. Kovatchev BP, Cox DJ, Gonder-Frederick LA, Clarke W. Symmetrization of the blood glucose measurement scale and its applications. *Diabetes Care* 1997 Nov;20(11):1655-1658. [doi: [10.2337/diacare.20.11.1655](https://doi.org/10.2337/diacare.20.11.1655)] [Medline: [9353603](https://pubmed.ncbi.nlm.nih.gov/9353603/)]
42. Hill NR, Hindmarsh PC, Stevens RJ, Stratton IM, Levy JC, Matthews DR. A method for assessing quality of control from glucose profiles. *Diabet Med* 2007 Jul;24(7):753-758. [doi: [10.1111/j.1464-5491.2007.02119.x](https://doi.org/10.1111/j.1464-5491.2007.02119.x)] [Medline: [17459094](https://pubmed.ncbi.nlm.nih.gov/17459094/)]
43. Zhu T, Uduku C, Li K, Herrero P, Oliver N, Georgiou P. Enhancing self-management in type 1 diabetes with wearables and deep learning. *NPJ Digit Med* 2022 Jun 27;5(1):78 [FREE Full text] [doi: [10.1038/s41746-022-00626-5](https://doi.org/10.1038/s41746-022-00626-5)] [Medline: [35760819](https://pubmed.ncbi.nlm.nih.gov/35760819/)]
44. R Core Team. *R: a language and environment for statistical computing*. R Foundation for Statistical Computing. 2019. URL: <https://www.r-project.org/> [accessed 2022-09-26]
45. Sarda-Espinosa A. dtwclust: time series clustering along with optimizations for the dynamic time warping distance. The Comprehensive R Archive Network. 2023 Feb 28. URL: <https://cran.r-project.org/web/packages/dtwclust/index.html> [accessed 2022-09-26]
46. FreeStyle Libre for glucose monitoring. National Institute for Health and Care Excellence. 2017 Jul 03. URL: <https://www.nice.org.uk/advice/mib110/resources/freestyle-libre-for-glucose-monitoring-pdf-2285963268047557> [accessed 2022-09-26]
47. Aleppo G, Ruedy KJ, Riddlesworth TD, Kruger DF, Peters AL, Hirsch I, REPLACE-BG Study Group. REPLACE-BG: a randomized trial comparing continuous glucose monitoring with and without routine blood glucose monitoring in adults with well-controlled type 1 diabetes. *Diabetes Care* 2017 Apr;40(4):538-545 [FREE Full text] [doi: [10.2337/dc16-2482](https://doi.org/10.2337/dc16-2482)] [Medline: [28209654](https://pubmed.ncbi.nlm.nih.gov/28209654/)]
48. Berndt DJ, Clifford J. Using dynamic time warping to find patterns in time series. In: *Proceedings of the 3rd International Conference on Knowledge Discovery and Data Mining*. 1994 Presented at: AAAIWS '94; July 31-August 1, 1994; Seattle, WA, USA p. 359-370.

49. Mazandarani FN, Mohebbi M. Wide complex tachycardia discrimination using dynamic time warping of ECG beats. *Comput Methods Programs Biomed* 2018 Oct;164:238-249. [doi: [10.1016/j.cmpb.2018.04.009](https://doi.org/10.1016/j.cmpb.2018.04.009)] [Medline: [29703454](https://pubmed.ncbi.nlm.nih.gov/29703454/)]
50. Skutkova H, Vitek M, Babula P, Kizek R, Provaznik I. Classification of genomic signals using dynamic time warping. *BMC Bioinformatics* 2013;14 Suppl 10(Suppl 10):S1 [FREE Full text] [doi: [10.1186/1471-2105-14-S10-S1](https://doi.org/10.1186/1471-2105-14-S10-S1)] [Medline: [24267034](https://pubmed.ncbi.nlm.nih.gov/24267034/)]
51. Govender P, Sivakumar V. Application of k-means and hierarchical clustering techniques for analysis of air pollution: a review (1980–2019). *Atmos Pollut Res* 2020 Jan;11(1):40-56 [FREE Full text] [doi: [10.1016/j.apr.2019.09.009](https://doi.org/10.1016/j.apr.2019.09.009)]
52. Edwards AW. Distances between populations on the basis of gene frequencies. *Biometrics* 1971 Dec;27(4):873-881. [Medline: [5138935](https://pubmed.ncbi.nlm.nih.gov/5138935/)]
53. English indices of deprivation 2019. National Statistics, UK. 2019. URL: <https://www.gov.uk/government/statistics/english-indices-of-deprivation-2019> [accessed 2022-09-26]
54. Jarrett RJ, Murrells TJ, Shipley MJ, Hall T. Screening blood glucose values: effects of season and time of day. *Diabetologia* 1984 Dec;27(6):574-577. [doi: [10.1007/BF00276970](https://doi.org/10.1007/BF00276970)] [Medline: [6543350](https://pubmed.ncbi.nlm.nih.gov/6543350/)]
55. Cappon G, Vettoretti M, Sparacino G, Facchinetti A. Continuous glucose monitoring sensors for diabetes management: a review of technologies and applications. *Diabetes Metab J* 2019 Aug;43(4):383-397 [FREE Full text] [doi: [10.4093/dmj.2019.0121](https://doi.org/10.4093/dmj.2019.0121)] [Medline: [31441246](https://pubmed.ncbi.nlm.nih.gov/31441246/)]
56. Engler R, Routh TL, Lucisano JY. Adoption barriers for continuous glucose monitoring and their potential reduction with a fully implanted system: results from patient preference surveys. *Clin Diabetes* 2018 Jan;36(1):50-58 [FREE Full text] [doi: [10.2337/cd17-0053](https://doi.org/10.2337/cd17-0053)] [Medline: [29382979](https://pubmed.ncbi.nlm.nih.gov/29382979/)]
57. Hossain MI, Yusof AF, Hussin AR, Iahad NA, Sadiq AS. Factors influencing adoption model of continuous glucose monitoring devices for internet of things healthcare. *Internet Things* 2021 Sep;15:100353 [FREE Full text] [doi: [10.1016/j.iot.2020.100353](https://doi.org/10.1016/j.iot.2020.100353)]

Abbreviations

BG: blood glucose
CGM: continuous glucose monitoring
COV: coefficient of variation
DM: diabetes mellitus
DTW: dynamic time warping
EPR: electronic patient record
FSL: FreeStyle Libre
GV: glycemic variability
HbA_{1c}: glycated hemoglobin
IMD: Index of Multiple Deprivation
NHS: National Health Service
T1DM: type 1 diabetes mellitus
TIR: time in range

Edited by K El Emam, B Malin; submitted 01.01.23; peer-reviewed by Y Jeem, P Wu; comments to author 02.02.23; revised version received 15.02.23; accepted 24.02.23; published 26.05.23.

Please cite as:

Chan NB, Li W, Aung T, Bazuaye E, Montero RM

Machine Learning–Based Time in Patterns for Blood Glucose Fluctuation Pattern Recognition in Type 1 Diabetes Management: Development and Validation Study

JMIR AI 2023;2:e45450

URL: <https://ai.jmir.org/2023/1/e45450>

doi: [10.2196/45450](https://doi.org/10.2196/45450)

PMID:

©Nicholas Berin Chan, Weizi Li, Theingi Aung, Eghosa Bazuaye, Rosa M Montero. Originally published in JMIR AI (<https://ai.jmir.org>), 26.05.2023. This is an open-access article distributed under the terms of the Creative Commons Attribution License (<https://creativecommons.org/licenses/by/4.0/>), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR AI, is properly cited. The complete bibliographic information, a link to the original publication on <https://www.ai.jmir.org/>, as well as this copyright and license information must be included.

Original Paper

An Assessment of How Clinicians and Staff Members Use a Diabetes Artificial Intelligence Prediction Tool: Mixed Methods Study

Winston R Liaw¹, MPH, MD; Yessenia Ramos Silva², BA; Erica G Soltero³, PhD; Alex Krist⁴, MPH, MD; Angela L Stotts⁵, PhD

¹Department of Health Systems and Population Health Sciences, Tilman J Fertitta Family College of Medicine, University of Houston, Houston, TX, United States

²Rice University, Houston, TX, United States

³USDA/ARS Children's Nutrition Research Center, Department of Pediatrics, Baylor College of Medicine, Houston, TX, United States

⁴Department of Family Medicine & Population Health, Virginia Commonwealth University School of Medicine, Richmond, VA, United States

⁵Department of Family & Community Medicine, McGovern Medical School, University of Texas Health Science Center at Houston, Houston, TX, United States

Corresponding Author:

Winston R Liaw, MPH, MD

Department of Health Systems and Population Health Sciences

Tilman J Fertitta Family College of Medicine

University of Houston

5055 Medical Circle

Houston, TX, 77204

United States

Phone: 1 713 743 9862

Email: winstonliaw@gmail.com

Abstract

Background: Nearly one-third of patients with diabetes are poorly controlled (hemoglobin A_{1c} ≥9%). Identifying at-risk individuals and providing them with effective treatment is an important strategy for preventing poor control.

Objective: This study aims to assess how clinicians and staff members would use a clinical decision support tool based on artificial intelligence (AI) and identify factors that affect adoption.

Methods: This was a mixed methods study that combined semistructured interviews and surveys to assess the perceived usefulness and ease of use, intent to use, and factors affecting tool adoption. We recruited clinicians and staff members from practices that manage diabetes. During the interviews, participants reviewed a sample electronic health record alert and were informed that the tool uses AI to identify those at high risk for poor control. Participants discussed how they would use the tool, whether it would contribute to care, and the factors affecting its implementation. In a survey, participants reported their demographics; rank-ordered factors influencing the adoption of the tool; and reported their perception of the tool's usefulness as well as their intent to use, ease of use, and organizational support for use. Qualitative data were analyzed using a thematic content analysis approach. We used descriptive statistics to report demographics and analyze the findings of the survey.

Results: In total, 22 individuals participated in the study. Two-thirds (14/22, 63%) of respondents were physicians. Overall, 36% (8/22) of respondents worked in academic health centers, whereas 27% (6/22) of respondents worked in federally qualified health centers. The interviews identified several themes: this tool has the potential to be useful because it provides information that is not currently available and can make care more efficient and effective; clinicians and staff members were concerned about how the tool affects patient-oriented outcomes and clinical workflows; adoption of the tool is dependent on its validation, transparency, actionability, and design and could be increased with changes to the interface and usability; and implementation would require buy-in and need to be tailored to the demands and resources of clinics and communities. Survey findings supported these themes, as 77% (17/22) of participants somewhat, moderately, or strongly agreed that they would use the tool, whereas these figures were 82% (18/22) for usefulness, 82% (18/22) for ease of use, and 68% (15/22) for clinic support. The 2 highest ranked factors affecting adoption were whether the tool improves health and the accuracy of the tool.

Conclusions: Most participants found the tool to be easy to use and useful, although they had concerns about alert fatigue, bias, and transparency. These data will be used to enhance the design of an AI tool.

(*JMIR AI* 2023;2:e45032) doi:[10.2196/45032](https://doi.org/10.2196/45032)

KEYWORDS

artificial intelligence; medical informatics; qualitative research; prediction tool; clinicians; diabetes; treatment; clinical decision support; decision-making; survey; interview; usefulness; implementation; validation; design; usability

Introduction

Background

Poor control, defined as a hemoglobin A_{1c} (HbA_{1c}) level >9.0%, contributes to complications, including nephropathy [1-6], retinopathy [4,7], and neuropathy [4,8]. Reducing poor control is important because a 2% decrease in HbA_{1c} (eg, from 9% to 7%) lowers the probability of microvascular complications by 50% to 76% [9]. The number of Americans with poorly controlled diabetes has been increasing, contributing to preventable morbidity and mortality [10-12]. In federally qualified health centers (FQHCs), the percentage with poor control was 32% in 2016 (up from 29% in 2009), suggesting that a new approach to diabetes management is needed [13,14]. Owing to the importance of poor control, the metric has been included in Healthy People 2030, which sets the national target at 11.6%, and in the measure sets that payers use to assess quality [15,16]. Thus, successfully reaching targets for diabetes control is important not only for patient health but also for the viability of health care organizations.

To meet these goals, researchers and clinicians are using artificial intelligence (AI) to integrate electronic health records (EHRs) and social risk factors, such as neighborhood characteristics, to predict outcomes important to individuals with diabetes, including poor control [17-22]. For instance, communities with poor housing, transportation, poverty, and education have higher rates of diabetes [23-25]. With the growth of EHRs, remote patient monitoring, and geo-tracking, the amount of data available to clinicians has increased exponentially [26]. Although this digitization offers tremendous opportunities for prediction, it also risks overwhelming clinicians [27]. This is true for primary care, which influences downstream spending and is responsible for whole person care that spans organs and diseases and serves as a point of integration with public health and behavioral health [28]. As a result of these functions, primary care clinicians are particularly susceptible to burnout, and it remains to be seen whether AI can help [29,30].

Unfortunately, the implementation of AI tools for diabetes has lagged, and few tools are used in practice, limiting their impact. A systematic review identified only 51 studies involving AI implementation [31]. Of these, 6 were related to diabetes. These applications used computer vision to diagnose diabetic retinopathy from retinal images and EHR data to predict those at risk for hyperglycemia. One study examined the implementation of a tool that predicts poor glycemic control [32]. As it was not tailored to the clinic's resources and population, only 14% (4/28) of users indicated that they would

recommend the tool to others, and many users reported that the interventions were inappropriate or not useful [32]. One possibility is that the organization failed to adequately address sociotechnical issues. The sociotechnical theory posits that the implementation of technology depends on values, mindsets, and communication and is an evolutionary process best achieved by early and active engagement with frontline workers [33,34]. Taken together, these studies indicate that a greater focus on AI implementation and end-user engagement during development are needed to tailor tools to clinical resources and workflows.

Objectives

As the absence of engagement has the potential to reduce trust and increase errors, researchers are starting to pay attention to end users [35] and are finding that usability of and satisfaction with AI tools are generally high [35-37]. Although most of these tools have targeted specialists, 1 study examined how primary care physicians use an AI tool to diagnose skin lesions [38]. Most of these studies used quantitative methods and examined tools that have already been developed [35-37]. This study is novel because it qualitatively assesses the use of a poorly controlled diabetes risk tool that has yet to be created and is based on the theory that early engagement with clinicians and staff will lead to methodological and design decisions that will support the tool's implementation. Furthermore, it is one of the few studies to target clinicians and staff working in primary care. The objective of this study was to assess how clinicians and staff would use and modify an AI clinical decision support tool for diabetes and to identify concerns and factors that affect its adoption and implementation.

Methods

Study Design and Participants

This is a mixed methods study of semistructured interviews and surveys to assess the perceived usefulness and ease of use, intent to use, and factors affecting tool adoption. The inclusion criteria were individuals (clinicians and staff) working in clinics that care for diabetes, adults aged ≥18 years, and English speakers. Participants were recruited via email through the researchers' networks.

Interview Procedures

Interviews were conducted by a trained interviewer between June 2021 and January 2022. All interviews were in English, conducted using a web-based platform, and audio recorded. Participants were compensated US \$50 upon completion of the interview and survey.

Ethics Approval

The protocol was approved by the Institutional Review Board of the University of Houston (STUDY00002980).

Interview Guide

A semistructured interview guide ([Multimedia Appendix 1](#)) was developed by the research team ([Textbox 1](#)). The questions were informed by the Technology Acceptance Model. This model was developed to predict individual adoption and use of new technology. It theorizes that individuals' intention to use new technology is determined by perceived usefulness, defined as "the extent to which a person believes that using [a new technology] will enhance his or her job performance," and perceived ease of use, defined as "the degree to which a person believes that using [a new technology] will be free of effort"

Textbox 1. Semistructured interview questions.

- What would you do with the information that you reviewed in the electronic health record alert?
- How useful, if at all, is this information for managing your patients with diabetes?
- What additional information would make this electronic health record alert more useful?
- How would you want the information presented to you so that it was easy to use?
- Would you prefer to receive this information at a specific point in time, such as at the point of care?
- To whom should this information be given? Consider clinicians, staff, administrators, and patients.
- What concerns do you have about using this tool?
- What are you already doing to identify people who are at high risk for uncontrolled diabetes?
- Besides uncontrolled diabetes, are there other undesirable outcomes that would be important to predict to improve the health of your patients?
- What are the factors that would affect whether this tool is implemented into practice at your clinic?

Qualitative Data Analysis

The interviews were transcribed using a web-based service (Otter [40]). A research assistant checked the transcripts for accuracy and cleaned and deidentified the transcripts when appropriate. The transcripts were coded by 2 individuals using thematic content analysis in NVivo (QSR International). First, the coders read each transcript independently. On the basis of the study objectives, interview guide, and responses, codes were generated using repeated ideas. Following the first reading, the coders compared the codes and developed a guiding codebook (version 1) with a list of codes and definitions. Using the updated codebook, the coders independently applied codes to the interviews in a second reading and met to reconcile coding discrepancies and modify the codebook (version 2). The coders used the resultant codebook to conduct a final review of the interviews, coming together to reconcile differences. Coding stopped once study objectives were saturated, indicating that no new information was identified. Following the coding process, codes were organized into themes and findings. To describe the strength of ideas, we calculated the number of respondents contributing to each finding.

Survey Design

Following the interview, the participants completed a survey ([Multimedia Appendix 2](#)). On a 7-point Likert scale (strongly disagree to strongly agree), participants reported their intent to

[39]. The model explains approximately 40% of the variance in individuals' intention to use a new technology and actual use [39]. During the interview, participants were asked to review a sample EHR alert and were informed that their clinic is considering the implementation of a clinical decision support tool that uses AI. This tool incorporates data from the EHR and the neighborhood in which the patient lives to predict whether the patient will have uncontrolled diabetes. The alert indicates that the fictional patient is at high risk for having an HbA_{1c} level of >9% over the next year. The tool suggests multiple actions that could reduce the risk, including sending referrals to a social worker, dietitian, or behavioral specialist, ordering an antidepressant or a diabetes medication, and scheduling visits every 3 months.

use the tool, perceived usefulness, ease of use, and organizational support for use. Next, they rank ordered the factors influencing the tool's implementation (cost of the tool, accuracy, health improvement, cost to the system, usability, impact on clinical workflows, and other). To quantify the extent to which AI would need to outperform clinician intuition for adoption, we asked participants to respond to the following prompt:

A team of clinicians and staff were tasked with predicting whether the 1000 individuals with diabetes in your practice would have a hemoglobin A_{1c} > 9% in the next year. The following year, your practice announced that the team accurately predicted the fate of 800 of these individuals. How many people would the AI tool need to accurately categorize for you to consider using it?

We collected demographic information, including age, gender, race and ethnicity, professional role, and practice setting. Physicians also reported the years since residency graduation and their specialty.

Quantitative Data Analysis

We used descriptive statistics to quantify demographics and responses.

Results

Overview

In total, 22 individuals participated in this study. They were predominantly women, Hispanic, and physicians (Table 1). The sample also included a nurse practitioner, physician assistant, behavioral therapist, and social worker. Overall, attitudes toward the tool were favorable (Table 2). Of 22 participants, 17 (77%) somewhat, moderately, or strongly agreed that they would use

the tool, whereas this figure was 18 (82%) for its usefulness. These figures were 82% (18/22) and 68% (15/22) for ease of use and clinic support, respectively. When asked to rank order the factors affecting implementation, the top 3 items were whether the tool improved health, accuracy, and usability. Finally, we asked participants to quantify how accurate the tool would need to be for them to consider using it. Of 1000 individuals with diabetes, the mean number of people whose prognosis the tool would need to accurately predict was 617 (SD 264), although the responses ranged from 20 to 900.

Table 1. Participant demographics (n=22).

Characteristics	Values, n (%)
Gender	
Women	13 (59)
Men	8 (36)
Prefer not to answer	1 (5)
Race and ethnicity (select all that apply)	
Hispanic, Latinx, or Spanish origin	9 (41)
White	6 (27)
Asian	4 (18)
Black or African American	1 (5)
Middle Eastern or North African	1 (5)
Prefer not to answer	1 (5)
Professional role	
Physician	14 (64)
Nurse practitioner	1 (5)
Physician assistant	1 (5)
Nurse	1 (5)
Behavioral specialist	1 (5)
Social worker	1 (5)
Other (front desk, administrative, or medical assistant)	3 (14)
Primary practice site	
Academic health center or faculty practice	8 (36)
Federally qualified health center or look-alike	6 (27)
Private solo or group practice	4 (18)
Health maintenance organization (eg, Kaiser Permanente)	2 (9)
Mental health center	1 (5)
Other (multiple sites)	1 (5)
Specialty (includes physicians, nurse practitioners, and physician assistants)	
Family medicine	15 (94)
Pediatrics	1 (6)
Years since residency graduation (physicians only)	
In residency	4 (29)
1-10	4 (29)
11-20	3 (21)
21-30	3 (21)

Table 2. Attitudes toward the tool and factors affecting implementation.

	Values
Attitudes^a, mean (SD)	1-7 ^b
“I would use the clinical decision support tool ^c .”	5.6 (1.4)
“I find the clinical decision support tool to be useful in my job.”	5.7 (1.3)
“I find the clinical decision support tool to be easy to use.”	5.8 (1.2)
“In general, the clinic would support my use of this clinical decision support tool.”	5.0 (1.7)
Factors affecting implementation (rank order)^d	
Factor, mean (SD)	
Whether its use improves health	2.5 (1.7)
Accuracy	2.7 (1.7)
Usability	3.7 (1.5)
Impact on clinic workflows	3.9 (1.6)
Cost	4.2 (1.7)
Whether its use reduces costs to the health care system	4.3 (1.6)
A team of clinicians and staff were tasked with predicting whether the 1000 individuals with diabetes in your practice would have a hemoglobin A_{1c} >9% in the next year. The following year, your practice announced that the team accurately predicted the fate of 800 of these individuals. How many people would the AI^e tool need to accurately categorize for you to consider using it?	
Values, mean (SD); range	617 (273); 20-900
Distribution of responses, n (%)	
0-200	3 (14)
201-400	1 (5)
401-600	6 (27)
601-800	6 (27)
801-1000	6 (27)

^a1 indicates strongly disagrees, and 7 indicates strongly agree.

^bRange of possible responses.

^cn=21.

^d1 indicates the most important factor, and 6 indicates the least important factor.

^eAI: artificial intelligence.

Multiple themes related to care delivery and concerns about the tool's use, adoption, and implementation emerged from the interviews (Table 3).

Table 3. Identified themes and subthemes (n=22).

Themes and subthemes	Participants, n (%)
How could the tool affect the delivery of care?	
This tool has the potential to be useful because it provides information that is not currently available and can make care more efficient and effective	
The tool is not currently available, addresses a clinical gap, and represents a departure from the status quo.	7 (32)
Clinicians and staff would increase their focus on diabetes, by scheduling more frequent visits, interacting with patients in between visits, managing diabetes even when acute issues emerge, and providing targeted education.	20 (91)
This tool could improve population health, address quality measures, and contribute to efficient resource allocation.	10 (45)
The tool would facilitate individualized and holistic care, by integrating primary care, behavioral health, and social care.	11 (50)
Participants were ambivalent about the tool's impact on populations that have been made susceptible. Some participants thought these were the patients who needed attention the most, whereas others thought that making a positive impact would be difficult.	7 (32)
What concerns do clinicians and staff have about the tool?	
Clinicians and staff were concerned about how the tool affects patient-oriented outcomes and clinic workflows	
Participants were concerned the tool would lead to harms, contribute to overdiagnosis, be used punitively, and make care more expensive.	15 (68)
The utility is limited for those clinicians who know their patients well or have access to existing programs, and some would rather focus on people who are already uncontrolled.	8 (36)
Participants were concerned that the tool would exacerbate existing problems, such as health disparities and alert fatigue.	14 (64)
Participants were concerned that the tool's accuracy and implementation were not supported by evidence.	5 (23)
What changes would increase adoption?	
Adoption of the tool is dependent on its validation, transparency, actionability, and design and could be increased with changes to the interface and usability	
The tool needs to be validated against patient-oriented outcomes so that clinics can quantify the potential return on their investment.	4 (18)
Knowing how the tool was developed and the rationale behind why an individual is high risk allows clinicians and staff to gauge the tool's credibility.	11 (50)
To act on the information, clinicians and staff need to understand which risk factors are modifiable and which actions will have the greatest impact on lowering risk.	6 (27)
Using user-centered design principles has the potential to minimize the tool's impact on workflows and maximize readability.	13 (59)
The ability to customize the tool is important because implementation could differ across practices and clinicians.	2 (9)
Participants recommended integrating functionality and relevant information from within the EHR ^a .	19 (86)
Participants recommended other events that could be predicted, including cardiovascular disease, uncontrolled hypertension, worsening depression, care gaps (eg, preventive services), and missed appointments.	22 (100)
What factors would affect implementation?	
Implementation would require buy-in and need to be tailored to the demands and resources of clinics and communities	
The local context affects what can be done in response to the information provided by the tool. Conversely, participants will become frustrated if the tool recommends an option that is not available.	12 (55)
Responding to the tool in a comprehensive manner requires the engagement of a comprehensive team. Although there was strong consensus regarding the role of clinicians and nurses, participants expressed ambivalence regarding administrators and patients.	21 (95)
Participants wanted to share this information with patients to empower them and support transparency but were also concerned that the information would cause confusion and stress.	20 (91)
There was a lack of consensus regarding when the alert should appear, with some wanting it at the point of care, whereas others wanted to review the information outside of visits (eg, periodic lists or a dashboard).	17 (77)
Successful implementation would require trialability, training, interoperability, and buy-in.	8 (36.)

^aEHR: electronic health record.

Theme 1

The tool has the potential to be useful because it provides information that is not currently available and can make care more efficient and effective.

When asked about how the tool could affect care, several participants (7/22, 32%) noted that such a tool does not exist and that it would fill a gap:

No, we don't already have a system. So I think there is value in adding a tool that would help improve care. [Physician, academic health center]

...a lot of it [clinician decisions] is...individual clinician suspicion...a lot of it is going to be based on how well each clinician knows their patients. [Physician, academic health center]

Other participants argued that the tool would facilitate the delivery of proactive care, building on the core function of primary care:

The primary argument for this tool...is that it's easier to prevent something than it is to cure it. [Physician, academic health center]

...the heart of what we do in primary care is to try to help patients with chronic conditions avoid long term complications of those conditions...if [AI believes] this person might be at greater risk, I might see [that patient] more often. I might spend more time with them. I might ask different questions because I would be trying to prevent [the complication]. [Physician, academic health center]

As a result of using the tool, clinicians and staff thought they would increase their focus on diabetes by scheduling more frequent visits, interacting with patients in between visits, managing diabetes even when acute issues emerge, and providing targeted education (20/22, 91%):

I find that for patients who are diabetic, it is the frequency of touches at every opportunity to control their diabetes that makes the biggest difference. And so if a patient has come in for a cold, or even anything else, other than diabetes, there's an opportunity to intervene. For those patients who are poorly controlled, it's usually because they're engaging with a system very infrequently. And so from that perspective, getting them reengaged in the system to become familiar with a system becomes the most valuable tool. [Physician, Health Maintenance Organization]

...it...makes you think twice...it...makes you pay attention a little bit closer, and makes [you] ask, okay, why are they at risk? What are the things that I can do to reduce the risk? [Physician, private solo or group practice]

...awareness is probably some of the best medicine you can give. And my philosophy is empowering a patient to give them the education, so they can make better decisions moving forward...I'm trying to

empower this patient to take control of their own care.

[Physician, private solo or group practice]

Others believed that the tool could be used to improve elements of population health, such as improving the quality of care delivered and allocating resources to high-need patients (10/22, 45%):

...as a clinician, it's part of my responsibility to have some awareness of the...health...of...my small population...And so this would help to do some of that. [Physician, private solo or group practice]

And also, it's part of our billing, and HEDIS measures anyway, we're supposed to have A1cs that are below eight, and so I feel like this is designed to meet that standard. [Physician, academic health center]

[Knowing which patients are at high risk is] kind of helpful...[it tells you] where to put your resources. [Nurse practitioner, FQHC]

By integrating information about mental health and social risk factors, our participants (11/22, 50%) believed that the tool would facilitate individualized, holistic care:

Now that [AI] has brought it up...I would explore things...that cause high A1c's like social determinants, depression, medical intensification... [Physician, academic health center]

I think it would be very useful, because it really takes a kind of a holistic approach of looking at the entire patient, and not just, I'm not just looking at like their blood sugar. [Behavioral specialist, FQHC]

I would provide education about the connection between depression and diabetes, and how they can very much go hand in hand, and how a diabetes diagnosis can either lead to a depression diagnosis or exacerbate depression that's already there. [Social worker, FQHC]

Participants were ambivalent about the tool's impact on susceptible populations. Some participants thought that these were the patients who needed attention the most, whereas others thought that making a positive impact would be difficult (7/22, 32%):

I think definitely...in [an] underserved population, it might be more beneficial, especially since they have less access to care. [Physician assistant, FQHC]

Say...I have...10 patients in the morning, and all of them have this alert, and so for all of them, I'm taking...these extra steps to identify barriers...that's going to take more of my time. [Physician, private solo or group practice]

The whole predicting, based on community or...based on where the person lives...struck me a little odd...it feels almost like...an overgeneralization...[because] you come from this community, you are at risk...Are we stereotyping?...Are we making assumptions...because someone comes from...poverty, or...a certain marginalized population? [Social worker, FQHC]

Theme 2

Clinicians and staff were concerned about how the tool affects patient-oriented outcomes and clinic workflows.

Participants had myriad concerns about the tool. First, they were concerned that the tool would lead to harms, contribute to overdiagnosis, be used punitively, and make care more expensive (15/22, 68%):

Would it make care worse? Yeah, potentially...So if you're prompted to prescribe medications...for people who are not yet at a certain level of risk, the [benefit to harm] ratio becomes smaller. [Physician, academic health center]

I would be concerned about [the] over identification [and] over diagnosis. [Physician, private solo or group practice]

I think that increasing the cost of care is definitely going to happen...in many systems because of how healthcare is paid for. So if I make a referral...for the patient, and the patient has to go and pay for the social worker [and] dietitian, I've just increased the cost of care. [Physician, academic health center]

I think that it is important to not make it look like...the fact that [patients are still uncontrolled]...is [because] you [are] a bad physician...I'm tired of that. [Physician, academic health center]

In particular, those clinicians who know their patients well or have access to existing programs thought the utility was limited, and some would rather focus on people who are already uncontrolled (8/22, 36%):

...a lot of it is going to be based on how well each clinician knows their patients, and how well and how comfortable the patient feels and speaking up on their own behalf for concerns that might have arisen. [Physician, academic health center]

We are asked on a monthly basis to review our patients who are not at a goal hemoglobin A1c level. Our...focus in the last six months has been...around Latino patients...So I find...this particular...information to be less valuable because we're kind of doing it on a monthly basis already. [Physician, health maintenance organization]

I would probably focus on the people I know who already have A1c's more than 9% and start working on that population first. [Physician, mental health center]

They were also concerned that the tool would exacerbate existing issues such as health disparities and alert fatigue:

Racial bias is...something that's implicitly existent in normal data sets...this is something that just compounds...It's like a small mistake that compounds into something bigger. [Physician, private solo or group practice]

...the primary concern stems from excess information being available...But if there's already a lot of data points, and they're not...actionable, it can be

overwhelming or just ignored. [Physician, health maintenance organization]

Finally, participants were concerned that the tool's accuracy and implementation would not be supported by evidence (5/22, 23%):

If it's things that are [inaccurate and] manually entered into the EHR system that are driving this..., it certainly could create false alerts and waste time or...miss people who actually are at risk because...things weren't...entered correctly, or left blank. [Physician, private solo or group practice]

You have to prove to me first that identifying and managing folks like this can actually help. [Physician, academic health center]

It's only useful if I trust the information. [Physician, academic health center]

Theme 3

Adoption of the tool is dependent on its validation, transparency, actionability, and design and could be increased by changing the interface and usability.

The tool needs to be validated against patient-oriented outcomes so that clinics can quantify the potential return on investment (4/22, 18%):

The factors would be how useful the tool is, first of all, how validated the tool is and if you can show that...it changes outcomes. [Physician, private solo or group practice]

The participants expected a degree of transparency and wanted to know how the tool was developed and the rationale behind the high risk of an individual. This information allows them to gauge the tool's credibility (11/22, 50%):

...if I'm going to use a tool, I want to be able to...click a link [that] will take me to the website and I can just learn more [about] where this is being trained. [Physician, private solo or group practice]

It would be helpful to know why that patient is at risk. And that will make you believe it or not. [Physician, private solo or group practice]

I think some sort of report that shows me which factors contributed the most to these alerts may help me even more. [Physician, academic health center]

Knowing why someone is at high risk is necessary but insufficient. Participants also wanted to understand which risk factors are modifiable and which actions will have the greatest impact on lowering risk (6/22, 27%):

...if the evidence says social work drops the risk by 50% [and] dietitian...drops the risk by 40%, on average, but in my patient, the alert fired because of nutritional concerns, I might choose the dietitian as a first choice because it might have a greater impact for this patient in particular. [Physician, academic health center]

...what would be really helpful...would be some sense of the potential impact of each of these, because I'm

not going to be able to get my patient to do all six potentially. But if they were organized in such a way to say this step will reduce the risk by this much. That step will reduce the risk by less...then I might be able to prioritize. [Physician, academic health center]

Participants believed that perceived usability and readability would be key drivers of adoption (13/22, 59%):

[Adoption] would depend very, very, very heavily on the provider perception of usefulness and usability. [Physician, academic health center]

Instead of showing six [actionable steps]...you...could [show] fewer options and color [code them]...from most benefit to least benefit. [Physician, academic health center]

The ability to customize the tool is important because implementation could differ across practices and clinicians (2/22, 9%):

...there's a lot of customization that would have to occur on the front end, to make sure that these...action items are clickable [and that] applicable resources [are] available. [Physician, private solo or group practice]

Participants recommended integrating functionality and relevant information within the EHR (19/22, 86%). They wanted to include a wide range of laboratories and vital signs to provide a context for risk prediction and broaden the types of actions that could be completed within the tool:

...one of the hard parts about managing diabetes is knowing...they need another agent, and then maybe which agent the insurance might cover...it would be even more beneficial if [the tool told] me these might be suggestive agents to add...for [better] control. [Physician, academic health center]

I'd want to know when and what their last hemoglobin A1c was and when their last appointment was. And then I want to know if they have seen a dietitian in the past and how long ago? [Physician, mental health center]

Participants thought that this model could be applied to other conditions and recommended that the tool be used to predict important events in primary care, including cardiovascular disease, uncontrolled hypertension, worsening depression, care gaps (eg, preventive services), and missed appointments (22/22, 100%):

...you could apply the same sort of thing to preventive care to any chronic disease to including depression, hypertension, coronary disease. [Physician, academic health center]

...how likely is this person going to follow through on their screenings, [like] getting their mammogram? [Physician, private solo or group practice]

Theme 4

Implementation would require buy-in and need to be tailored to the demands and resources of clinics and communities.

The local context affects what can be performed in response to the information provided by the tool. Conversely, participants will become frustrated if the tool recommends an option that is not available (12/22, 55%):

[My use of the tool] would depend a great deal on what resources are actually available to me. [Physician, academic health center]

...depending on...what your clinics resources are, if you're getting alerts for people that you have no ability to help, because you don't have access to a social worker...that doesn't feel really good. [Physician, academic health center]

Responding to the tool in a comprehensive manner requires the engagement of a comprehensive team. Although there was strong consensus regarding the role of clinicians and nurses, participants expressed ambivalence regarding administrators and patients (21/22, 95%). All members of the primary care team have potential roles to play, including front desk personnel, pharmacists, and social workers. As roles differ for each practice, the recipients of the information may be practice dependent:

...staff should have the means to be able to respond to...this...there would be a lot of a lot of value in having multiple eyes on this to make sure that nobody falls through the cracks. [Physician, in residency]

I don't think this would be terribly helpful for administrators. Sometimes it's used punitively. And I don't think that that's what we want. [Physician, academic health center]

Regarding who should receive this information: "I feel like each location might want to designate that." [Physician, academic health center]

Participants wanted to share this information with patients to empower them and support transparency but were also concerned that the information would cause confusion and stress (20/22, 91%). They thought that the information without context could be harmful and that they would need scripts to explain the results in a patient-centric manner:

I think [who should receive the information] would be very, very practice dependent...I think giving the information to patients can be really valuable. I think how it's presented and how it's framed [is important]. [Physician, academic health center]

I think just a lack of context for the patient on why these certain things were ordered would be [a] concern for high alert with the patient...[patients] having no clue what it means could create...panic or some distress in the patient. [Physician, in residency]

There was a lack of consensus regarding when the alert should appear, with some wanting it at the point of care, whereas others wanted to review the information outside of visits (eg, periodic lists or a dashboard, 17/22, 77%):

This really depends on the operator. For me...if it comes too early, I'll lose it...So...I feel like [the timing] should be adjustable. That would be best

because every provider is very different. [Physician, private solo or group practice]

Another thing would be making sure that it's the right time. So again, if I'm in room with the patient, personally, I don't want to see these pop up, because I'm probably goal-oriented at that moment where I'm trying to put in something specific and this would just slow me down. [Physician, academic health center]

I would be more likely to address it...if it was something I was prompted with when I opened the labs specifically...I'm going there to review their hemoglobin...I'm going there to review their lipids...so if I'm going [to the chart] for that, and...I'm prompted with this, then then I'm going to be more likely to address it right at that moment. [Physician, in residency]

I wouldn't want a list of 500 patients, because there's no way that anybody's going to keep track of that...that would be very difficult. [Social worker, FQHC]

Successful implementation would require trialability, training, interoperability, and buy-in (8/22, 36%):

I would definitely be open to trialing it but would do it in a quality improvement sort of a mindset where we saw how things were going beforehand and how things were going afterwards. And if it didn't help me, then I wouldn't continue using it. [Physician, private solo or group practice]

Also takes education. So educating providers about what this alert is and what this means and what we do with it. [Social worker, FQHC]

Discussion

Principal Findings

From the surveys, respondents found the tool to be useful and easy to use and, if available, would use it. During the interviews, they noted that the tool is not available now and would generally change their behavior. With notable exceptions, many participants reported that their organizations lacked a systematic approach for reducing the percentage of those who are poorly controlled. Despite these benefits, the tool was not uniformly accepted, with several respondents indicating that it did not provide useful information for those patients who are well known to the practice and for those practices already offering comprehensive services. Others were concerned that AI would perpetuate biases and that alert fatigue would contribute to burnout. To enhance adoption, respondents wanted to know why the patients were at risk and what could be done to reduce that risk. Finally, they wanted to be able to tailor the tool to their local environment, noting that the suggestions offered and

the recipients of the information needed to be customized to the resources, needs, and workflows of their unique clinics.

Our findings align with, and build on, the work of others. For example, similar to our results, other clinicians have responded favorably to the usability of tools that use AI [36,37]. Although usability and accuracy were deemed important, our respondents asked for steps that could be taken in response to predictions and wanted to know that those actions would lead to better health, echoing the sentiment found in other studies [35]. Similar to others, they also regarded the technology with skepticism [35,41,42]. For many years, researchers and policy makers have issued warnings regarding the black-box nature of AI and its role in widening disparities [43,44]. Our findings demonstrate that these are not theoretical issues. The clinicians and staff members in our study called for greater explainability (ie, justifications for the tool's output), wanted these issues explicitly acknowledged and addressed, and cautioned that these tools will continue to languish on shelves in the absence of satisfactory solutions [44]. They are concerned about how AI can perpetuate the racial biases embedded within data sets and about their role in supporting biased systems. Taken together, these findings highlight the importance of the tool's actionability, explainability, and harm minimization (resulting from bias and workflow disruptions) for its implementation and provide a blueprint for researchers interested in developing AI tools for primary care settings. For example, to address these concerns, researchers must engage communities and end users early in the development process to identify and mitigate sources of bias and iteratively test and refine the tool's impact [45].

There are several limitations to this study that should be considered when interpreting these results. First, because we recruited participants from our networks, many of them were from academic settings and FQHCs. Our results may differ if we had a sample that is more representative of primary care clinics across the United States. Second, we did not ask the participants to use a prototype of the tool when responding to the questions. If they had, their responses to the questions regarding ease of use and usefulness may have been different. However, we contend that incorporating input from end users before a prototype is created is important for adoption. Finally, we did not assess other factors that influence adoption, such as computer self-efficacy, that we did not assess.

Conclusions

Most participants found the tool to be easy to use and useful. They also believed that the tool could improve population health and contribute to individualized care. Conversely, participants were concerned about alert fatigue, bias, and transparency. To gauge the tool's credibility, they wanted to know why the patients were at high risk and what they could do to reduce that risk. These data will be used to inform the development of an AI tool for diabetes.

Acknowledgments

This study was supported by a grant from the American Board of Family Medicine Foundation (principal investigator: WRL).

Conflicts of Interest

WRL received funding from the American Board of Family Medicine Foundation.

Multimedia Appendix 1

Interview guide.

[[DOCX File , 84 KB - ai_v2i1e45032_app1.docx](#)]

Multimedia Appendix 2

Clinician survey.

[[DOCX File , 18 KB - ai_v2i1e45032_app2.docx](#)]

References

1. ADVANCE Collaborative Group, Patel A, MacMahon S, Chalmers J, Neal B, Billot L, et al. Intensive blood glucose control and vascular outcomes in patients with type 2 diabetes. *N Engl J Med* 2008 Jun 12;358(24):2560-2572 [[FREE Full text](#)] [doi: [10.1056/NEJMoa0802987](#)] [Medline: [18539916](#)]
2. -. Effects of intensive glucose lowering in type 2 diabetes. *N Engl J Med* 2008 Jun 12;358(24):2545-2559. [doi: [10.1056/nejmoa0802743](#)]
3. Duckworth W, Abraira C, Moritz T, Reda D, Emanuele N, Reaven PD, et al. Glucose control and vascular complications in veterans with type 2 diabetes. *N Engl J Med* 2009 Jan 08;360(2):129-139. [doi: [10.1056/nejmoa0808431](#)]
4. Fullerton B, Jeitler K, Seitz M, Horvath K, Berghold A, Siebenhofer A. Intensive glucose control versus conventional glucose control for type 1 diabetes mellitus. *Cochrane Database Syst Rev* 2014 Feb 14;2014(2):CD009122 [[FREE Full text](#)] [doi: [10.1002/14651858.CD009122.pub2](#)] [Medline: [24526393](#)]
5. The Diabetes Control and Complications (DCCT) Research Group. Effect of intensive therapy on the development and progression of diabetic nephropathy in the Diabetes Control and Complications Trial. The Diabetes Control and Complications (DCCT) Research Group. *Kidney Int* 1995 Jun;47(6):1703-1720 [[FREE Full text](#)] [doi: [10.1038/ki.1995.236](#)] [Medline: [7643540](#)]
6. Coca SG, Ismail-Beigi F, Haq N, Krumholz HM, Parikh CR. Role of intensive glucose control in development of renal end points in type 2 diabetes mellitus: systematic review and meta-analysis intensive glucose control in type 2 diabetes. *Arch Intern Med* 2012 May 28;172(10):761-769 [[FREE Full text](#)] [doi: [10.1001/archinternmed.2011.2230](#)] [Medline: [22636820](#)]
7. The Diabetes Control and Complications Trial Research Group. The relationship of glycemic exposure (HbA1c) to the risk of development and progression of retinopathy in the diabetes control and complications trial. *Diabetes* 1995;44(8):968-983. [doi: [10.2337/diab.44.8.968](#)]
8. -. The effect of intensive diabetes therapy on the development and progression of neuropathy. The Diabetes Control and Complications Trial Research Group. *Ann Intern Med* 1995 Apr 15;122(8):561-568. [doi: [10.7326/0003-4819-122-8-199504150-00001](#)] [Medline: [7887548](#)]
9. American Diabetes Association. Glycemic targets: standards of medical care in diabetes—2020. *Diab Care* 2020;43(Supplement 1):S66-S76. [doi: [10.2337/dc20-s006](#)]
10. Comprehensive Diabetes Care (CDC). NCQA. URL: <http://www.ncqa.org/report-cards/health-plans/state-of-health-care-quality/2016-table-of-contents/diabetes-care> [accessed 2017-11-29]
11. Quality: Traditional MIPS Requirements. Quality Payment Program. URL: <https://qpp.cms.gov/mips/quality-measures> [accessed 2017-11-29]
12. Crowley MJ, Holleman R, Klammer ML, Bosworth HB, Edelman D, Heisler M. Factors associated with persistent poorly controlled diabetes mellitus: clues to improving management in patients with resistant poor control. *Chronic Illn* 2014 Dec 24;10(4):291-302 [[FREE Full text](#)] [doi: [10.1177/1742395314523653](#)] [Medline: [24567193](#)]
13. Shi L, Lebrun LA, Zhu J, Hayashi AS, Sharma R, Daly CA, et al. Clinical quality performance in U.S. health centers. *Health Serv Res* 2012 Dec 17;47(6):2225-2249 [[FREE Full text](#)] [doi: [10.1111/j.1475-6773.2012.01418.x](#)] [Medline: [22594465](#)]
14. National Health Center Program Uniform Data System (UDS) Awardee Data. Health Resources & Services Administration. URL: <https://bphc.hrsa.gov/uds/datacenter.aspx#fn8> [accessed 2018-02-22]
15. Reduce the proportion of adults with diabetes who have an A1c value above 9 percent. Office of Disease Prevention and Health Promotion. URL: <https://tinyurl.com/dkx26uwx> [accessed 2022-09-06]
16. Diabetes: Hemoglobin A1c (HbA1c) Poor Control (> 9%). Centers for Medicare & Medicaid Services. URL: https://qpp.cms.gov/docs/ecqm-specs/2017/EC_CMS122v5_NQF0059_Diab_HbA1c_Ctrl/CMS122v5.html [accessed 2021-08-13]
17. Basu S, Narayanaswamy R. A prediction model for uncontrolled type 2 diabetes mellitus incorporating area-level social determinants of health. *Med Care* 2019 Aug;57(8):592-600. [doi: [10.1097/MLR.0000000000001147](#)] [Medline: [31268954](#)]
18. Basu S, Sussman JB, Berkowitz SA, Hayward RA, Yudkin JS. Development and validation of Risk Equations for Complications Of type 2 Diabetes (RECODE) using individual participant data from randomised trials. *Lancet Diabetes Endocrinol* 2017 Oct;5(10):788-798. [doi: [10.1016/s2213-8587\(17\)30221-8](#)]

19. Luo G. Automatically explaining machine learning prediction results: a demonstration on type 2 diabetes risk prediction. *Health Inf Sci Syst* 2016 Mar 8;4(1):2 [FREE Full text] [doi: [10.1186/s13755-016-0015-4](https://doi.org/10.1186/s13755-016-0015-4)] [Medline: [26958341](https://pubmed.ncbi.nlm.nih.gov/26958341/)]
20. Devireddy L, Dunn D, Sherman M. A feature-selection based approach for the detection of diabetes in electronic health record data. Michael W Sherman. 2014 May 7. URL: <http://www.michaelwsherman.com/projects/diabetes/report.pdf> [accessed 2023-05-17]
21. Berkman LF, Kawachi I, Glymour MM, editors. *Social Epidemiology* (2 edition). Oxford, England: Oxford University Press; Jul 2014.
22. Krieger N, Chen JT, Waterman PD, Rehkopf DH, Subramanian SV. Race/ethnicity, gender, and monitoring socioeconomic gradients in health: a comparison of area-based socioeconomic measures--the public health disparities geocoding project. *Am J Public Health* 2003 Oct;93(10):1655-1671. [doi: [10.2105/ajph.93.10.1655](https://doi.org/10.2105/ajph.93.10.1655)] [Medline: [14534218](https://pubmed.ncbi.nlm.nih.gov/14534218/)]
23. Cox M, Boyle PJ, Davey PG, Feng Z, Morris AD. Locality deprivation and Type 2 diabetes incidence: a local test of relative inequalities. *Soc Sci Med* 2007 Nov;65(9):1953-1964. [doi: [10.1016/j.socscimed.2007.05.043](https://doi.org/10.1016/j.socscimed.2007.05.043)] [Medline: [17719709](https://pubmed.ncbi.nlm.nih.gov/17719709/)]
24. Butler DC, Petterson S, Phillips RL, Bazemore AW. Measures of social deprivation that predict health care access and need within a rational area of primary care service delivery. *Health Serv Res* 2013 Apr 20;48(2 Pt 1):539-559 [FREE Full text] [doi: [10.1111/j.1475-6773.2012.01449.x](https://doi.org/10.1111/j.1475-6773.2012.01449.x)] [Medline: [22816561](https://pubmed.ncbi.nlm.nih.gov/22816561/)]
25. Sundquist K, Eriksson U, Mezuk B, Ohlsson H. Neighborhood walkability, deprivation and incidence of type 2 diabetes: a population-based study on 512,061 Swedish adults. *Health Place* 2015 Jan;31:24-30 [FREE Full text] [doi: [10.1016/j.healthplace.2014.10.011](https://doi.org/10.1016/j.healthplace.2014.10.011)] [Medline: [25463914](https://pubmed.ncbi.nlm.nih.gov/25463914/)]
26. Culbertson N. The skyrocketing volume of healthcare data makes privacy imperative. *Forbes*. 2021 Aug 6. URL: <https://www.forbes.com/sites/forbestechcouncil/2021/08/06/the-skyrocketing-volume-of-healthcare-data-makes-privacy-imperative/> [accessed 2022-08-29]
27. Medscape lifestyle report 2016: bias and burnout. Medscape. URL: <https://www.medscape.com/slideshow/lifestyle-2016-overview-6007335> [accessed 2021-03-22]
28. Friedberg MW, Hussey PS, Schneider EC. Primary care: a critical review of the evidence on quality and costs of health care. *Health Aff (Millwood)* 2010 May;29(5):766-772. [doi: [10.1377/hlthaff.2010.0025](https://doi.org/10.1377/hlthaff.2010.0025)] [Medline: [20439859](https://pubmed.ncbi.nlm.nih.gov/20439859/)]
29. Shanafelt TD, West CP, Sinsky C, Trockel M, Tutty M, Satele DV, et al. Changes in burnout and satisfaction with work-life integration in physicians and the general US working population between 2011 and 2017. *Mayo Clin Proc* 2019 Sep;94(9):1681-1694 [FREE Full text] [doi: [10.1016/j.mayocp.2018.10.023](https://doi.org/10.1016/j.mayocp.2018.10.023)] [Medline: [30803733](https://pubmed.ncbi.nlm.nih.gov/30803733/)]
30. Muehlematter UJ, Daniore P, Vokinger KN. Approval of artificial intelligence and machine learning-based medical devices in the USA and Europe (2015–20): a comparative analysis. *Lancet Digital Health* 2021 Mar;3(3):e195-e203. [doi: [10.1016/s2589-7500\(20\)30292-2](https://doi.org/10.1016/s2589-7500(20)30292-2)]
31. Yin J, Ngiam KY, Teo HH. Role of artificial intelligence applications in real-life clinical practice: systematic review. *J Med Internet Res* 2021 Apr 22;23(4):e25759 [FREE Full text] [doi: [10.2196/25759](https://doi.org/10.2196/25759)] [Medline: [33885365](https://pubmed.ncbi.nlm.nih.gov/33885365/)]
32. Romero-Brufau S, Wyatt KD, Boyum P, Mickelson M, Moore M, Cognetta-Rieke C. A lesson in implementation: a pre-post study of providers' experience with artificial intelligence-based clinical decision support. *Int J Med Inform* 2020 May;137:104072. [doi: [10.1016/j.ijmedinf.2019.104072](https://doi.org/10.1016/j.ijmedinf.2019.104072)] [Medline: [32200295](https://pubmed.ncbi.nlm.nih.gov/32200295/)]
33. Cherns A. Principles of sociotechnical design revisited. *Human Relation* 2016 Apr 22;40(3):153-161. [doi: [10.1177/001872678704000303](https://doi.org/10.1177/001872678704000303)]
34. Greenhalgh T, Wherton J, Papoutsis C, Lynch J, Hughes G, A'Court C, et al. Beyond adoption: a new framework for theorizing and evaluating nonadoption, abandonment, and challenges to the scale-up, spread, and sustainability of health and care technologies. *J Med Internet Res* 2017 Nov 01;19(11):e367 [FREE Full text] [doi: [10.2196/jmir.8775](https://doi.org/10.2196/jmir.8775)] [Medline: [29092808](https://pubmed.ncbi.nlm.nih.gov/29092808/)]
35. Asan O, Choudhury A. Research trends in artificial intelligence applications in human factors health care: mapping review. *JMIR Hum Factors* 2021 Jun 18;8(2):e28236 [FREE Full text] [doi: [10.2196/28236](https://doi.org/10.2196/28236)] [Medline: [34142968](https://pubmed.ncbi.nlm.nih.gov/34142968/)]
36. Milne-Ives M, de Cock C, Lim E, Shehadeh MH, de Pennington N, Mole G, et al. The effectiveness of artificial intelligence conversational agents in health care: systematic review. *J Med Internet Res* 2020 Oct 22;22(10):e20346 [FREE Full text] [doi: [10.2196/20346](https://doi.org/10.2196/20346)] [Medline: [33090118](https://pubmed.ncbi.nlm.nih.gov/33090118/)]
37. Abdulaal A, Patel A, Al-Hindawi A, Charani E, Alqahtani SA, Davies GW, et al. Clinical utility and functionality of an artificial intelligence-based app to predict mortality in COVID-19: mixed methods analysis. *JMIR Form Res* 2021 Jul 28;5(7):e27992 [FREE Full text] [doi: [10.2196/27992](https://doi.org/10.2196/27992)] [Medline: [34115603](https://pubmed.ncbi.nlm.nih.gov/34115603/)]
38. Micocci M, Borsci S, Thakerar V, Walne S, Manshadi Y, Edridge F, et al. Attitudes towards trusting artificial intelligence insights and factors to prevent the passive adherence of GPs: a pilot study. *J Clin Med* 2021 Jul 14;10(14):3101 [FREE Full text] [doi: [10.3390/jcm10143101](https://doi.org/10.3390/jcm10143101)] [Medline: [34300267](https://pubmed.ncbi.nlm.nih.gov/34300267/)]
39. Venkatesh V, Bala H. Technology acceptance model 3 and a research agenda on interventions. *Decision Sci* 2008 May;39(2):273-315. [doi: [10.1111/j.1540-5915.2008.00192.x](https://doi.org/10.1111/j.1540-5915.2008.00192.x)]
40. Otter.ai homepage. Otter.ai. URL: <https://otter.ai/> [accessed 2023-05-17]
41. Strohm L, Hehakaya C, Ranschaert ER, Boon WP, Moors EH. Implementation of artificial intelligence (AI) applications in radiology: hindering and facilitating factors. *Eur Radiol* 2020 Oct 26;30(10):5525-5532 [FREE Full text] [doi: [10.1007/s00330-020-06946-y](https://doi.org/10.1007/s00330-020-06946-y)] [Medline: [32458173](https://pubmed.ncbi.nlm.nih.gov/32458173/)]

42. Scott IA, Carter SM, Coiera E. Exploring stakeholder attitudes towards AI in clinical practice. *BMJ Health Care Inform* 2021 Dec 09;28(1):e100450 [FREE Full text] [doi: [10.1136/bmjhci-2021-100450](https://doi.org/10.1136/bmjhci-2021-100450)] [Medline: [34887331](https://pubmed.ncbi.nlm.nih.gov/34887331/)]
43. Obermeyer Z, Powers B, Vogeli C, Mullainathan S. Dissecting racial bias in an algorithm used to manage the health of populations. *Science* 2019 Oct 25;366(6464):447-453. [doi: [10.1126/science.aax2342](https://doi.org/10.1126/science.aax2342)] [Medline: [31649194](https://pubmed.ncbi.nlm.nih.gov/31649194/)]
44. Fjeld J, Achten N, Hilligoss H, Nagy A, Srikumar M. Principled artificial intelligence: mapping consensus in ethical and rights-based approaches to principles for AI. Berkman Klein Center for Internet & Society. 2020. URL: <https://dash.harvard.edu/handle/1/42160420> [accessed 2023-05-17]
45. Harmon D, Carter R, Cohen-Shelly M, Svatikova A, Adedinsewo DA, Noseworthy PA, et al. Real-world performance, long-term efficacy, and absence of bias in the artificial intelligence enhanced electrocardiogram to detect left ventricular systolic dysfunction. *Eur Heart J Digit Health* 2022 Jun;3(2):238-244 [FREE Full text] [doi: [10.1093/ehjdh/ztac028](https://doi.org/10.1093/ehjdh/ztac028)] [Medline: [36247412](https://pubmed.ncbi.nlm.nih.gov/36247412/)]

Abbreviations

AI: artificial intelligence

EHR: electronic health record

FQHC: federally qualified health center

HbA_{1c}: hemoglobin A_{1c}

Edited by K El Emam, B Malin; submitted 13.12.22; peer-reviewed by H Wang, S Sharma; comments to author 06.01.23; revised version received 09.03.23; accepted 22.04.23; published 29.05.23.

Please cite as:

Liaw WR, Ramos Silva Y, Soltero EG, Krist A, Stotts AL

An Assessment of How Clinicians and Staff Members Use a Diabetes Artificial Intelligence Prediction Tool: Mixed Methods Study
JMIR AI 2023;2:e45032

URL: <https://ai.jmir.org/2023/1/e45032>

doi: [10.2196/45032](https://doi.org/10.2196/45032)

PMID: [38875578](https://pubmed.ncbi.nlm.nih.gov/38875578/)

©Winston R Liaw, Yessenia Ramos Silva, Erica G Soltero, Alex Krist, Angela L Stotts. Originally published in JMIR AI (<https://ai.jmir.org>), 29.05.2023. This is an open-access article distributed under the terms of the Creative Commons Attribution License (<https://creativecommons.org/licenses/by/4.0/>), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR AI, is properly cited. The complete bibliographic information, a link to the original publication on <https://www.ai.jmir.org/>, as well as this copyright and license information must be included.

Original Paper

Identifying the Question Similarity of Regulatory Documents in the Pharmaceutical Industry by Using the Recognizing Question Entailment System: Evaluation Study

Nidhi Saraswat^{1*}, MSc; Chuqin Li^{1*}, PhD; Min Jiang¹, PhD

Eli Lilly and Company, Indianapolis, IN, United States

*these authors contributed equally

Corresponding Author:

Min Jiang, PhD

Eli Lilly and Company

893 Delaware St

Indianapolis, IN, 46225

United States

Phone: 1 615 926 8277

Email: jiang_min@lilly.com

Abstract

Background: The regulatory affairs (RA) division in a pharmaceutical establishment is the point of contact between regulatory authorities and pharmaceutical companies. They are delegated the crucial and strenuous task of extracting and summarizing relevant information in the most meticulous manner from various search systems. An artificial intelligence (AI)-based intelligent search system that can significantly bring down the manual efforts in the existing processes of the RA department while maintaining and improving the quality of final outcomes is desirable. We proposed a “frequently asked questions” component and its utility in an AI-based intelligent search system in this paper. The scenario is further complicated by the lack of publicly available relevant data sets in the RA domain to train the machine learning models that can facilitate cognitive search systems for regulatory authorities.

Objective: In this study, we aimed to use AI-based intelligent computational models to automatically recognize semantically similar question pairs in the RA domain and evaluate the Recognizing Question Entailment-based system.

Methods: We used transfer learning techniques and experimented with transformer-based models pretrained on corpora collected from different resources, such as Bidirectional Encoder Representations from Transformers (BERT), Clinical BERT, BioBERT, and BlueBERT. We used a manually labeled data set that contained 150 question pairs in the pharmaceutical regulatory domain to evaluate the performance of our model.

Results: The Clinical BERT model performed better than other domain-specific BERT-based models in identifying question similarity from the RA domain. The BERT model had the best ability to learn domain-specific knowledge with transfer learning, which reached the best performance when fine-tuned with sufficient clinical domain question pairs. The top-performing model achieved an accuracy of 90.66% on the test set.

Conclusions: This study demonstrates the possibility of using pretrained language models to recognize question similarity in the pharmaceutical regulatory domain. Transformer-based models that are pretrained on clinical notes perform better than models pretrained on biomedical text in recognizing the question’s semantic similarity in this domain. We also discuss the challenges of using data augmentation techniques to address the lack of relevant data in this domain. The results of our experiment indicated that increasing the number of training samples using back translation and entity replacement did not enhance the model’s performance. This lack of improvement may be attributed to the intricate and specialized nature of texts in the regulatory domain. Our work provides the foundation for further studies that apply state-of-the-art linguistic models to regulatory documents in the pharmaceutical industry.

(JMIR AI 2023;2:e43483) doi:[10.2196/43483](https://doi.org/10.2196/43483)

KEYWORDS

regulatory affairs; frequently asked questions; FAQs; Recognizing Question Entailment system; RQE system; transformer-based models; textual data augmentations

Introduction

Regulatory Affairs

In a pharmaceutical company, the regulatory affairs (RA) department is responsible for obtaining approval for new pharmaceutical products and ensuring that approval is maintained for as long as the company wants to keep the product on the market. It serves as the interface between the regulatory authorities (such as the Food and Drug Administration, European Medicines Agency, etc) and pharmaceutical companies. It is the responsibility of the RA department to keep abreast of current legislation, guidelines, and other regulatory intelligence.

Regulatory data sources are dynamic and enormous. Regulatory professionals go through the extremely tedious and grueling task of extracting relevant information for various regulatory tasks. The process includes generating one or more suitable key phrases; searching for these key phrases in multiple data sources; and combining appropriate information retrieved from different data sources into a clear, compact, and concise summary of findings. Keeping track of such large data sources for relevant information manually is difficult and complex. An artificial intelligence-powered search system can drastically reduce manual efforts and improve the efficiency and quality of the existing processes.

The question answering (QA) system is an efficient approach for retrieving information. Much research has been conducted on open-domain QA systems based on deep learning techniques owing to the availability of vast data sources. However, the medical domain received less attention owing to the shortage of medical data sets. Although electronic health records empower the field of medical QA by providing medical information to answer user questions, the gap remains significant in the medical domain, especially for text-based sources.

The intricate challenges of automated QA in the biomedical domain are growing with the increasing diversity and specialization of medical texts. One of the promising tracks investigated in QA is to map new questions to formerly answered questions that are “similar.” Frequently asked questions (FAQ) component in an intelligent search system can considerably speed up the automated search system and enhance the status of search results. Therefore, an FAQ model component that interacts with the user query input to return a similar question that has already been asked in the recent past can significantly accelerate the remaining components of the search system pipeline and improve the system’s effectiveness.

In this study, we proposed a new approach for detecting similar questions based on Recognizing Question Entailment (RQE) in the RA domain. We considered FAQs as a valuable and widespread source of information.

RQE is a crucial component of modern QA systems. The RQE approach for a QA system is to retrieve answers to a given question proposed by users using natural languages by retrieving answers to an entailed and already answered question. The answered question and its associated answer are saved in a question-answer pair database. Question entailment is formally defined by Ben Abacha and Demner-Fushman [1] as follows:

question A entails question B if every answer to question B is also a correct answer to question A exactly or partially. It is a challenging task to understand questions and judge the semantic similarity of two questions: (1) one question could be rephrased in many different ways and (2) two different questions may refer to the same problem and could be answered by the same answer [2].

Background

RQE in the General Domain

Researchers in the general domain used 2 public benchmark data sets for question similarity tasks: SemEval and Quora question pair. These 2 data sets have labeled training data for question-question similarity. SemEval-2017 task 3 [3] featured questions from subforums of StackExchange, a family of technical community support forums. Quora question pair data set contains pairs of similar questions asked by people on the Quora website. The topics of these questions range from philosophy to entertainment. The best-performing systems for the SemEval question similarity task used syntactic tree kernels or the SoftCosine metric [4]. Kunneman et al [5] compared 2 recent approaches (SoftCosine and Smoothed partial tree kernel) and 2 traditional approaches (BM25 [6] and translation-based language model) and showed that the choice of a preprocessing method and a word-similarity metric have a considerable impact on the final result. Shah et al [7] first applied the adversarial domain adaptation to the problem of duplicate question detection across different domains and outperformed the best baseline on StackExchange questions. More recently, Nguyen et al [8] outperformed previous studies on the SemEval data set by combining a convolutional neural network and features from external knowledge to measure the similarity between 2 questions. In addition to these studies on the aforementioned popular data sets, Wang et al [9] used a method based on the Coattention-DenseGRU (gated recurrent unit) to match similar questions on Chinese rice-related questions.

Although many researchers have put efforts into recognizing general question similarity, their approaches do not generalize well to domains that require domain expert knowledge, such as the biomedical domain. First, questions in the biomedical domain demand much domain-specific knowledge, and a single word can change the meaning of the question [10]. Second, there are few publicly available biomedical question-question similarity data sets, resulting in a limited number of samples that can be used to train models that can effectively learn those differences. Given the increasing popularity of RQE-based QA systems, question similarity in the biomedical domain is currently an active research area. A growing number of RQE-based QA systems have been proposed, and an international challenge was held in 2019 [11].

RQE in Biomedical Domain

A wide range of approaches has been proposed to capture the semantic relationship between pairs of questions in RQE-based QA systems. Luo et al [12] calculated similarities between questions using statistical syntactic features and Unified Medical Language System annotated semantic features. Ben Abacha and Demner-Fushman [1] used machine learning models with lexical

features and semantic features to determine the similarity of question pairs. More recent studies have gone beyond traditional feature-based methods and used deep learning models. Wang and Nyberg [13] used a dual entailment approach with bidirectional recurrent neural networks and attention mechanisms to predict question similarity. Ben Abacha and Demner-Fushman [14] improved their system using feature-based logistic regression and neural network that passed the concatenated sentence representations to multiple ReLU layers to classify question pairs into entailment or no entailment categories. McCreery et al [10] augmented a general language model with medical knowledge by using a double fine-tuning process. A pretrained language model is first fine-tuned with a large general corpus (eg, Quora question pairs) and then fine-tuned with a small number of labeled question pairs.

The MEDIQA 2019 challenge [11] included 3 tasks: natural language interface (NLI), RQE, and QA in the medical domain. It aimed to further research efforts to improve domain-specific information retrieval and question-answer systems. In the challenge, approaches using ensemble methods and transfer learning of multitask language models outperformed traditional deep learning models for RQE task [11]. The PANLP team [15] achieved the best result on RQE task by fine-tuning the pretrained language models, Bidirectional Encoder Representations from Transformers (BERT) [16] and multitask-deep neural network (DNN) [17]. They further boosted the performance on the RQE task by transfer learning from the NLI task. The Sieg team [18] ranked second for RQE tasks and used a multitask learning approach, with shared layers trained for the NLI on the RQE task. Approaches that used ensemble methods without multitask language models [19] ranked third in the competition, and approaches that used multitask models without ensemble methods [20] ranked fourth. More recently, Sarrouiti et al [21] proposed a multitask transfer learning method based on data augmentation for RQE. They outperformed other teams on the RQE test set of the 2019 MEDIQA challenges.

RQE or similarity is part of another more general natural language processing (NLP) task called semantic textual similarity (STS). Tasks of STS include comparing 2 sentences, 2 paragraphs, or even 2 documents. RQE is more closely related to QA and information retrieval systems.

STS in the General Domain

STS is connected to textual entailment (TE) and paraphrasing; however, it differs in many ways and is more directly applicable to several NLP tasks. Semantic similarity or STS is a task in NLP that scores the relationship between texts or documents using a defined metric. The aim is to identify the likeness or similarity in the meaning of 2 pieces of text.

STS differs from TE in that it assumes bidirectional graded equivalence between a pair of textual snippets. In the case of TE, the equivalence is directional; for example, a car is a vehicle, but a vehicle is not necessarily a car. STS also differs from both TE and paraphrasing in that rather than being a binary yes-or-no decision (eg, a vehicle is not a car), we defined STS to be a graded similarity notion (eg, a vehicle and a car are more similar than a wave and a car).

STS in the Biomedical Domain

STS in the clinical domain can empower stakeholders to detect and eliminate redundant information that may reduce the cognitive burden and improve the clinical decision-making process. The description in the study by Wang et al [22] discusses the details of the task of identifying clinical STS (ClinicalSTS). The participating systems were asked to return a numerical score, ranging from 0 to 5, indicating the degree of semantic similarity between the pair of 2 clinical sentences. The performance was measured using the Pearson correlation coefficient between the predicted similarity scores and human judgments.

1. The winning team submitted 4 systems. The first system was the random forest model using 63 features including string similarity features, entity similarity features, number similarity features, and deep learning features. The second system used the average score of the first system and dense neural networks. The third system, which was also the best-performing system among all submitted systems with a Pearson correlation of 0.8328, applied a regression model on 8 trained models including the random forest model, the Bayesian Ridge regression model, the Lasso regression model, the linear regression model, the Extra Tree model, the DNN using the Universal Sentence Encoder, the DNN using the inferSent encoder, and the Encoder-multilayer perceptron using the inferSent encoder. The fourth system used the average score of the first system, and the Encoder-multilayer perceptron used the inferSent encoder.
2. The team that placed second in this challenge used attention-based convolutional neural network (ABCNN) and bidirectional long short-term memory (Bi-LSTM) networks. One of their submissions used ABCNN with traditional NLP features. The second is a hybrid model of ABCNN and Bi-LSTM, with traditional NLP features. The third run ensembled the previous 2 systems by calculating the average scores. The ensemble model performed the best among their submitted systems.
3. The third-placed team proposed a sentence-embedding method that represents a sentence as a weighted average of word vectors, followed by a soft projection. They used a self-regularized identity map named Conceptors to correct the common component bias in linear sentence embedding. Majority voting and 2 different support vector regression models with only word embedding representation features were explored by the fourth-placed team for their submissions. The best performance was achieved by the majority voting method.

Lastra-Díaz and García-Serrano [23] presented an empirical study on the impact of a number of model design choices on a BERT-based approach to clinical STS. It was demonstrated that the proposed hierarchical convolution mechanism outperformed several alternative conventional pooling methods. Different parameter fine-tuning strategies with varying degrees of flexibility were investigated, and the optimal number of trainable transformer blocks was identified, thereby preventing overtuning. Finally, the utility of 2 data augmentation methods (segment reordering and back translation) on clinical STS was verified.

Hadj Taieb et al [24] proposed a novel framework based on a gated network to fuse distributed representation and one-hot representation of sentence pairs. Some state-of-the-art distributed representation methods, including convolutional neural network, Bi-LSTM, and BERT, were used in this framework, and a system based on this framework was developed for a shared task regarding clinical STS organized by BioCreative and OHNLP in 2018.

Elavarasi et al [25] demonstrated transformer-based models (BERT, XLNet, and RoBERTa) and developed a system that can use various transformer algorithms for measuring clinical STS. STS system has two modules: (1) a transformer model-based feature learning module that learns distributed sentence-level representations from sentence pairs and (2) a regression-based similarity score learning module that calculates similarity score between 0 and 5 according to the distributed representations derived from the transformers. The authors explored several methods to combine the distributed representations from different transformers, including (1) simple head-to-tail concatenation, (2) pooling, and (3) convolution. The experiment's results showed that the RoBERTa model achieved the best performance compared with other transformer models.

The work done in the study by Lastra-Díaz et al [26] focuses on ranking the degree of similarity between clinical texts. The paper studied the impact of using different preprocessing methods as well as different feature representation methods (word embeddings–BioWordVec vs sentence embeddings–BioSentVec) by proposing a system with a simple neural network. The study demonstrated that sentence embeddings provided superior text representation than word embeddings, better capturing sentence semantics, whereas word embeddings were not a distant performer. It was observed that word embeddings benefited from using a more thorough text-preprocessing pipeline, whereas sentence embeddings obtained better test results with a basic preprocessing approach.

Data Sets for STS

This subsection briefly describes some of the popular data sets at the sentence pairs level that are used to evaluate the semantic similarity algorithms. The performance of various semantic similarity algorithms is measured by the correlation of the achieved results with that of the standard measures available in these data sets. Li et al [27] used a data set comprises 65 sentence pairs that were created using the dictionary definition of 65 word pairs used in the Rubenstein-Goodenough data set [28]. A similarity range of 0 to 4 (from the lowest to the highest) was provided voluntarily by 32 native English speakers. The mean of the scores given by all the volunteers was taken as the final score. The SICK data set [29] consists of 10,000 sentence pairs derived from 2 existing data sets, the ImageFlickr 8 and MSR-Video descriptions data sets. Each sentence pair is associated with a relatedness score and a text entailment relation. The relatedness score ranges from 1 to 5, and the 3 entailment relations are “NEUTRAL, ENTAILMENT, and CONTRADICTION.” The annotation was performed using crowdsourcing techniques. The STS [30-34] data sets were built by combining sentence pairs from different sources by the

organizers of the SemEval shared task. The data set was annotated using Amazon Mechanical Turk and verified by the organizers themselves. Various sources such as newswire, videos, glosses, Workshop on Machine Translation evaluation, Machine Translation evaluation, newswire headlines, forum posts, news summary, image descriptions, tweet news pairs, student answers, QA forum answers, and committed belief were used to build the STS data set.

The computation of semantic similarity between various types of text fragments such as words, sentences, or documents plays a key role in a wide range of NLP tasks such as information retrieval [35], text summarization [36], text classification [37], essay evaluation [38], machine translation [39], and QA [40,41].

A wide range of semantic similarity measures has been proposed and applied in various applications and domains. These measures vary in performance based on their approaches and application domains. Detailed comparisons of these measures can be found in previous work [22,42-47].

Amir et al [42] proposed a semantic similarity algorithm using kernel functions. They used constituency-based tree kernels where the sentence is broken down into subject, verb, and object based on the assumption that most semantic properties of a sentence are attributed to these components. The input sentences are parsed using the Stanford Parser to extract various combinations of subject, verb, and object. The similarity between the various components of the given sentences is calculated using a knowledge base, and different averaging techniques are used to average the similarity values to estimate the overall similarity, and the best among them is chosen based on the root mean squared error value for a particular data set. Benedetti et al [43] proposed a novel knowledge-based technique, Context Semantic Analysis, for estimating interdocument similarity. The technique is based on a Semantic Context Vector, which can be extracted from a knowledge base and stored as metadata of a document and used to compute interdocument similarity. The authors also demonstrated how Context Semantic Analysis can be effectively applied in the information retrieval domain, even if user queries, typically composed of a few words, contain a limited number of entities. Yang et al [44] presented a response prediction model that learns a sentence encoder from conversations. The encoder learned from the input-response pairs performs well on sentence-level STS. The basic conversation model learned from Reddit conversations is competitive with existing sentence-level encoders on public STS tasks. A multitask model trained on Reddit and Stanford NLI classification achieved the state-of-the-art for sentence encoding-based models on the STS Benchmark task. An FAQ retrieval system with a method using query-question similarity and BERT-based query answer relevance was proposed by Sakata et al [48]. A traditional unsupervised information retrieval system is used to calculate the similarity between the query and the question. In contrast, the relevance between the query and answer, calculated using BERT model, are learned using QA pairs in an FAQ database. Minaee and Liu [49] evaluated the proposed approach on two data sets: (1) localgovFAQ, a data set that is constructed in a Japanese administrative municipality domain, and (2) StackExchange data set, which is the public data set in English.

Uva et al [50] proposed to inject structural relationships in neural networks by (1) learning a support vector machine model using tree kernels on relatively few pairs of questions (a few thousands), as gold standard training data are typically scarce; (2) predicting labels on a very large corpus of question pairs; and (3) pretraining neural networks on such a large corpus. The experiments in the study were performed on the Quora and SemEval question similarity data sets. A deep learning-based model for automatic QA was proposed [51] to solve the use case of customer case service automation. The questions and answers are embedded using neural probabilistic modeling (doc2vec), followed by training a deep similarity neural network to determine the similarity score of a pair of answer and question. For each question, the best answer is found as the one with the highest similarity score. Cai et al [51] trained this model on a large-scale public QA database and then fine-tuned it to transfer to the customer care chat data.

About Transfer Learning

The advancements of deep learning in NLP in recent years have improved, accelerated, and automated various functions and features of text analytics. Deep learning enables models to understand and learn the meaning of words and phrases in different language contexts. However, all these utilities demand large and complex deep learning models that are data hungry. They require training with thousands or millions of data points before making a plausible prediction. Training is expensive in terms of both time and resources. The issue with such models is that they are performed only on a single task. Future tasks require a new set of data points and a greater number of resources. Transfer learning comes into the picture by transferring knowledge learned from one model to another.

Transfer learning is a machine learning method where a model trained on one task is repurposed on a second related task as an optimization that allows rapid progress when modeling the second task. It can train DNNs with comparatively fewer data.

We subsequently briefly describe a few DNN models experimented with in this paper that use the transfer learning approach.

BERT

Google's BERT [16] has significantly altered the NLP landscape in recent years. BERT is a contextualized word representation model based on a masked language model and pretrained using bidirectional transformers. It is designed to pretrain deep bidirectional representations from the unlabeled text by jointly conditioning on both the left and right context. As a result, the pretrained BERT model can be fine-tuned with only one additional output layer to create state-of-the-art models for a wide range of NLP tasks.

BERT is pretrained on a large corpus of unlabeled text, including the entire Wikipedia (2500 million words) and Book Corpus (800 million words). BERT is a "deeply bidirectional" model, meaning that BERT learns information from both the left and the right side of a token's context during the training phase.

BERT architecture builds on top of the transformer. All these transformer layers are encoder-only blocks. BERT is pretrained on 2 NLP tasks: masked language modeling and next-sentence prediction. The pretrained BERT has a maximum of 512 input tokens (position embeddings). The output would be a vector for each input token. Each vector is composed of 768 float numbers (hidden units).

Clinical BERT

BERT model is pretrained in general text corpora. A specific model pretrained on specialty corpora, such as clinical text, is available in the form of Clinical BERT, a modified BERT model. Specifically, the representations are learned using medical notes and further processed for downstream clinical tasks. Clinical BERT [52] models are pretrained on 2 types of data: one for generic clinical text and another for discharge summaries. Similar to BERT, Clinical BERT is a trained transformer encoder stack. Clinical BERT is also a bidirectional transformer.

BioBERT

BioBERT [53] is a domain-specific language representation model that is pretrained on large-scale biomedical corpora. BioBERT is specifically pretrained on PubMed abstracts (PubMed) and PubMed Central full-text articles along with English Wikipedia and Book Corpus data sets as in BERT.

BlueBERT

The success of the General Language Understanding Evaluation, which was primarily to help the development of pretrained language models based on performance on generic NLP tasks, led to the development of Biomedical Language Understanding Evaluation (BLUE). BLUE is similar to General Language Understanding Evaluation but is more specific to the biomedical domain. The benchmark consists of 5 tasks with 10 data sets covering biomedical and clinical texts with different data set sizes and difficulties. BlueBERT [54], which was originally named National Center for Biotechnology Information BERT, was pretrained on PubMed abstracts and MIMIC-III (Medical Information Mart for Intensive Care) clinical notes. The work done by Peng et al [54] focused on experimenting BLUE in conjunction with Embeddings from Language Model and BERT models. BlueBERT was found to be the best-performing model and significantly superior to other models in the clinical domain.

Table 1 summarizes the pretraining details of different BERT models used in the experiments of this study.

Table 1. Summary of pretraining details for the various Bidirectional Encoder Representations from Transformers (BERT) models used in our experiments.

Model	Vocabulary	Pretraining	Corpus	Text size
BERT	Wikipedia+Books	N/A ^a	Wikipedia+Books	3.3B words (16 GB)
Clinical BERT	Wikipedia+Books	Continual pretraining	MIMIC ^b (subset)+MIMIC-III	0.5B words (3.7 GB)
BioBERT	Wikipedia+Books	Continual pretraining	PubMed+PMC ^c	4.5B words
BlueBERT	Wikipedia+Books	Continual pretraining	PubMed+MIMIC-III	4.5B words

^aN/A: not applicable.

^bMIMIC: Medical Information Mart for Intensive Care.

^cPMC: PubMed Central.

Methods

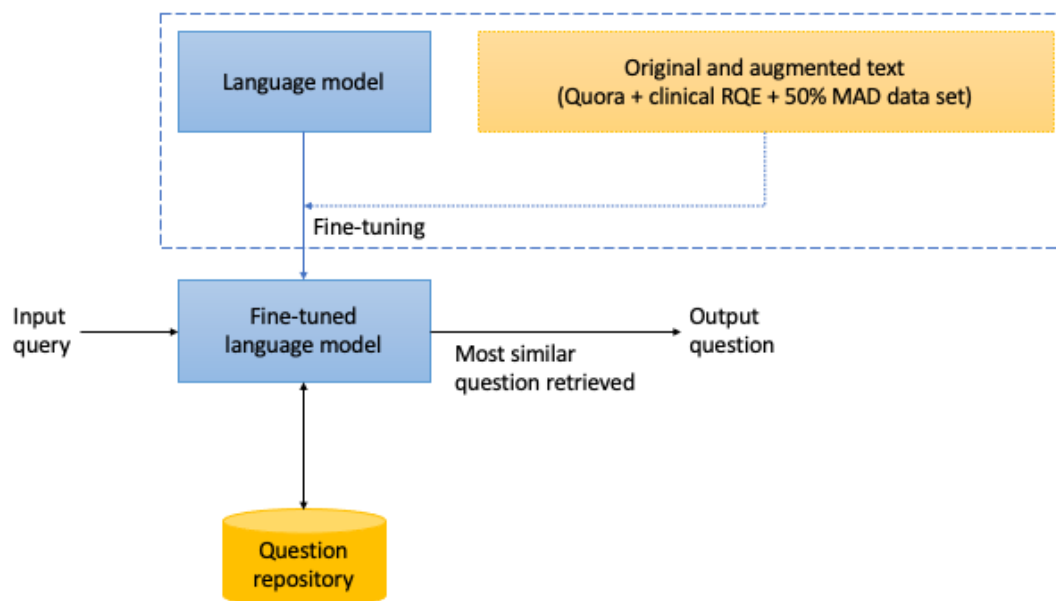
DNN Architecture

Figure 1 describes the working of our proposed FAQ system. The proposed FAQ system uses 2 major components: a question repository and a fine-tuned language model. The FAQ repository, which acted as the source of questions to identify entailment or no entailment for input queries, was maintained in the proposed FAQ system. The input query to the fine-tuned

language model was compared against each question in the question repository to identify and retrieve the most similar FAQ, if any.

The language model was fine-tuned by using the Quora question pairs and clinical RQE (C-RQE) data sets. Different experimental and data split strategies were used to identify the best-performing model configuration. These data sets and experimental strategies are explained in detail in the following subsections.

Figure 1. Architecture diagram for frequently asked questions (FAQ) system. MAD: manually annotated data set; RQE: Recognizing Question Entailment.



Data Sets

The experiments in this paper were based on these three different data sets.

1. *Quora questions pairs (Quora)*: The Quora question pairs data set [55] provides an opportunity to train and test models of semantic equivalence, based on actual Quora data. Each line in the data set contains an ID for each question in the question pair, a unique ID for the question pair, the full text for each question, and a binary label that indicates whether the line contains a duplicate question pair. Table 2 presents a few sample lines of the data set. This data set contains over 400,000 labeled question pairs. Of the 404,290

question pairs, 255,027 (63.08%) had a negative (0) label and 149,263 (36.92%) had a positive (1) label, making our data set unbalanced.

2. *C-RQE*: The work done in the study by Ben Abacha and Demner-Fushman [1] describes an automatic method for constructing training corpora for RQE. The RQE data set constructed in this paper used the National Library of Medicine collection of 4655 clinical questions asked by family physicians. The resulting C-RQE data set had approximately 8588 question pairs in the form of an XML, with RQE value labels as true or false.
3. *Regulatory RQE—manually annotated data set (MAD)*: The subject matter experts, who are part of the organizational RA team, manually annotated a collection

of 268 question pairs with entailment and no entailment labels. Of these 268 question pairs, 127 were entailment pairs and 141 were no entailment pairs. The records in this

data set were of the following format: (question_pair_ID, label, question1, question2). Some of the example records from this data set are presented in Table 3.

Table 2. Samples of Quora question pairs.

ID	Question1 ID	Question2 ID	Question1	Question2
447	895	896	What are natural numbers	What is a least natural number?
1518	3037	3038	Which pizzas are the most popularly ordered pizzas on Domino's menu?	How many calories does a Domino's pizza have?
3272	6542	6543	How do you start a bakery?	How can one start a bakery business?
3362	6722	6723	Should I learn python or Java first?	If I had to choose between learning Java and Python, what should I choose to learn first?

Table 3. Samples of question pairs in the test set.

ID	Label	Question 1	Question 2
1	Entailment	Is Emend approved by the European Medicines Agency?	Has the European Medicines Agency authorized Emend?
2	Entailment	What is Emend approved indication in the European Union?	What is the indication of Emend in the European Union product information?
15	Not_entailment	Does Tisagenlecleucel has gotten an orphan designation by the European Medicines Agency?	Does Tisagenlecleucel refused by the European Medicines Agency?
16	Not_entailment	Is ELIANA (other IDs: NCT02435849/CCTL019B2202) a single arm trial?	Is ELIANA (other IDs: NCT02435849/CCTL019B2202) a randomized arm trial?

Preprocessing of Data Sets

Both the Quora and C-RQE data sets were transformed to a format that was consistent with the MAD data set. The Quora data set was transformed to this format by removing individual question IDs and converting "is_duplicate" binary field to "entailment/no entailment" label field ("is_duplicate=1" indicates entailment label and vice versa). In contrast, the C-RQE data set, which is an XML, was converted to the format consistent with MAD by extracting ID, question1, question2, and value labels. The "value label=true" was transformed into an entailment label and vice versa.

Data Split Strategy

Three data sets are commonly used in deep learning model development: training, validation, and test sets. The model is trained on the training set, and the validation set is used to evaluate the model fit unbiasedly during the hyperparameter tuning stage. The test set is independent from the training and validation sets and is used to assess the model's performance.

The experiments designed in this study are built on 2 types of data set split strategy described as follows:

Strategy 1: Quora and C-RQE data sets were used as training and validation sets, respectively. With this strategy, we have 404,283 sentence pairs in the training data set, 7143 pairs in the validation data set, and 150 pairs in the testing data set.

Strategy 2: Quora and C-RQE data sets were combined to further split them into training and validation sets such that the training set had approximately 90% of the records, whereas the remaining 10% were part of the validation set. Therefore, the training set had 90% of the records from Quora and C-RQE data sets. The validation set comprised 10% of the records from

Quora and C-RQE data sets, which were not part of the training set. The validation set also included 50% of MAD, which was not part of the test set. With this strategy, we have 370,282 sentence pairs in the training data set, 41,279 pairs in the validation data set, and 150 pairs in the testing data set.

Model Evaluation

The variation of experiments conducted in this study for the Quora, C-RQE, and MAD data sets were performed on top of 4 types of BERT models: (1) regular BERT [16], (2) Clinical BERT [52], (3) BioBERT [53], and (4) BlueBERT [54]. The performance of several types of model configurations was evaluated for accuracy, F_1 -score metrics, and the area under the receiver operating characteristic curve (AUC). The model's accuracy was estimated by finding the total number of true entailment or no entailment predictions out of the total number of predictions done by the model. F_1 -score was an error metric that was calculated from the precision and recall of the test. F_1 -score of the model was interpreted as the harmonic mean of the precision and recall, conveying the balance between the precision and recall of the model.



F_1 -score ranges from 0 to 1, with 0 being the worst and 1 being the best value. The highest value of 1 indicates that the model has a perfect precision and recall, whereas the lowest value of 0 indicates that either the precision or recall is 0.

AUC is another commonly used statistical metric that evaluates the performance of classification models and provides a comprehensive measure of a model's ability to classify instances from different classes correctly. The AUC metric has advantages

over accuracy and F_1 -score in that it is insensitive to data imbalance and considers the model's behavior across all possible classification thresholds. The AUC value lies in the range of 0 to 1, where a higher value indicates a more robust ability for classification. An AUC of ≤ 0.5 suggests a model with no predictive power other than random guessing.

Experimental Design

The experiments performed in this paper used regular BERT, Clinical BERT, BioBERT, and BlueBERT models, which were fine-tuned on the Quora, C-RQE, and MAD data sets. Each of these models was fine-tuned on the 2 data split strategies. For experimentation of regular BERT, we used bert-base-uncased, whereas the Clinical BERT model used in this paper was pretrained on clinical notes. The BlueBERT-Base, Uncased, PubMed+MIMIC-III variant of the BlueBERT model was experimented with in this paper.

We used the Hugging face transformers library for model fine-tuning and text classification. The following set of hyperparameters was established to be the best set of

hyperparameters and was used for all the experiments conducted in this paper: epochs=3, learning rate= 3×10^{-5} , batch size=32, and maximum sequence length=150.

Ethical Considerations

This study is not human participant research; thus, no ethics approval was sought.

Results

The experimental design of this paper is described in [Table 4](#).

[Table 4](#) describes the performance of different BERT models on the datasets discussed in the *Data Split Strategy* section. As baseline experiments, we used experiments 1, 4, 7, and 10 to assess the performance of the models without transferring any prior knowledge. With comparable accuracy, F_1 -score, and AUC, the Clinical BERT and BioBERT models outperformed the other 2 baseline models, whereas the BlueBERT model performed the lowest with an accuracy of 48.9%, F_1 -score of 0.328, and AUC of 0.242.

Table 4. Performance of different Bidirectional Encoder Representations from Transformers (BERT) models on data sets: without augmentation.

Experiment	Model	Data split strategy	Accuracy (%)	F_1 -score	AUC ^a
1	BERT	N/A ^b	48.9	0.328	0.303
2	BERT	1	82	0.808	0.920
3	BERT	2	90.66	0.904	0.958
4	Clinical BERT	N/A	58.6	0.586	0.584
5	Clinical BERT	1	90	0.894	0.971
6	Clinical BERT	2	90	0.897	0.961
7	BioBERT	N/A	54.1	0.513	0.612
8	BioBERT	1	66	0.538	0.729
9	BioBERT	2	56.66	0.515	0.580
10	BlueBERT	N/A	48.9	0.328	0.242
11	BlueBERT	1	82	0.807	0.920
12	BlueBERT	2	84.66	0.842	0.920

^aAUC: area under the receiver operating characteristic curve.

^bN/A: not applicable.

Regardless of the data split approach, the performance of all models enhanced after being fine-tuned with domain-specific data. BERT's accuracy, F_1 -score, and AUC improved the most after being fine-tuned with data split strategy 1. The accuracy of the model increased from 48.9% to 90.66%. The performance of BioBERT model showed minimal improvement. The accuracy of the BioBERT model increased from 54.1% to 66% after being fine-tuned with our data split strategy 1. Although the accuracy of BioBERT model improved from 54.1% to 56.66%, the model's classification capability decreased because AUC decreased from 0.612 to 0.580.

The best-performing models were BERT (data split strategy 2) and Clinical BERT (data split strategy 1 and 2) with an accuracy of 90.66%, 90%, and 90%; F_1 -score values of 0.904, 0.894, and 0.897; and AUC of 0.958, 0.971, and 0.961, respectively.

Experiments 1 and 2 used the general BERT model to provide 82% and 90.66% accuracy for data split strategy 1 and 2, respectively. This behavior of data split strategy 2 surpassing data split strategy 1 was consistent across all BERT models experimented in [Table 4](#), except for the BioBERT model. The Clinical BERT model with both data split strategies was among the top-performing models with an accuracy of approximately 90% and an AUC of >0.96 . The BioBERT model did not fare very well compared with all the other models in [Table 4](#), with an accuracy of 66% and 56.66% for data split strategy 1 and 2, respectively. The BlueBERT model performed noticeably better than BioBERT, with an accuracy of approximately 82% and 84.66% for data split strategy 1 and 2, respectively, which was still lower than that of the high-performing BERT and Clinical BERT models.

Discussion

Principal Findings

In this study, we used computational models to recognize question entailment in pharmaceutical regulatory domains. As there is no publicly available labeled data set in this field, we adopted the idea of transfer learning. We fine-tuned 4 different versions of pretrained BERT language models on 2 publicly available data sets (Quora and C-RQE). The best model achieved 90.66% accuracy in RQE on our MAD, which contained 150 question pairs in the regulatory field. To the best of our knowledge, this study is the first to use state-of-the-art NLP models to recognize question semantic similarity in the pharmaceutical regulatory domain. Our study could provide the foundation for future studies that apply NLP technologies to text in the pharmaceutical regulatory domain.

As shown in [Table 4](#), the BERT model outperformed the other BERT variants in terms of its ability to learn domain knowledge using transfer learning. Although the BERT model performed poorly on the test data set before fine-tuning, its accuracy increased in RQE after being fine-tuned using domain-specific question pairs. This finding was also supported by experiments 2 and 3. The model based on BERT did not perform well with our data split strategy 1 (experiment 2). However, it reached the highest accuracy when we fine-tuned the BERT with our data split strategy 2 (experiment 3). Our data split strategy 1 used only Quora's general domain question pairs as training resources. In contrast, strategy 2 includes both general domain question pairs and clinical questions from C-RQE as part of the training and validation sets. This indicates that BERT can perform well in the pharmaceutical regulatory domain text if we provide sufficient clinical domain background knowledge to the model and fine-tune it.

We also found that Clinical BERT models outperformed other BERT variants in this specific domain. Clinical narratives from general and nonclinical biomedical text have known differences in linguistic characteristics [52]. All BERT variants used in this study were initialized from BERT, but they were pretrained on the corpus from different fields. The Clinical BERT model was pretrained with clinical notes, the BioBERT model was pretrained with biomedical corpus, and the BlueBERT model was pretrained with the combination of biomedical text and clinical notes. We found that the Clinical BERT and BlueBERT models performed better than the BioBERT model. In other words, the models that were pretrained with clinical notes from MIMIC-III data set have a better performance than the models pretrained with PubMed articles in our RQE task. A possible reason is that the nature of questions in the regulatory domain, shown in [Table 3](#), resonates more closely with the clinical notes text genre. This finding highlights the importance of pretraining with the proper text genre in learning the context-dependent representation [54].

Although DNNs perform well in a variety of NLP tasks, a large number of data are required to train deep learning models. The lack of training data has become one of the significant challenges to training deep learning models in the biomedical field, which could lead to underfitting models and could reduce

their performance. We do not have a publicly available labeled data set for the pharmaceutical regulatory domain. Instead, we fine-tuned pretrained language models on the C-RQE data set to learn domain-specific knowledge. In our previous experiments, only 21% of the question pairs in the training corpus were from the regulatory-related domain. Consequently, we extended our experiments by expanding our training data set with data augmentation technologies. We aimed to study the impact and utility of augmentation techniques on pharmaceutical domain text using the general BERT and Clinical BERT models.

Researchers in the field of computer vision commonly use data augmentation to expand the number and variety of data without collecting new data. They create new image samples by rotating, changing the color, cropping, and compressing the images. Unlike images, the data in NLP are discrete, making it more challenging to generate high-quality augmented examples efficiently and effectively in the field of NLP.

With the increasing interest in and demand for data augmentation in NLP, many text data augmentation technologies have been proposed. Back translation is the most popular data augmentation method. The back translation approach involves translating a sequence into another language and then back to the original language. Deep learning models, such as Seq2Seq [56], neural machine translation [57], and transformers [58], can be used to translate. Various rule-based techniques have also been used in data augmentation. Wei and Zou [59] proposed Easy Data Augmentation, including synonym replacement, random insertion, deletion, and swapping. For paraphrase identification, Chen et al [60] built a signed graph over the data, with each sentence as nodes and labels as edges. They used balance theory and transitivity to induce augmented sentence pairs based on the graph. Kang et al [61] extended the Easy Data Augmentation method for biomedical named entity recognition by incorporating the Unified Medical Language System knowledge. Another class of techniques uses multiple samples to generate new pieces, pioneered by MixUp [62], which interpolates the inputs and labels of ≥ 2 examples. The difficult part of using MixUp in NLP is that it requires a continuous input. This issue was overcome by Chen et al [63], who mixed embeddings or higher layers. Some other model-based approaches used the text generation models, such as GPT-2 [64], to generate candidate examples from the training data set. Some trade-offs should be considered when choosing from these methods.

We used 2 data augmentation techniques in this study, entity replacement and back translation. The entity replacement technique in this study used the Scispacy [65] named entity recognition model trained on the BC5CDR corpus to identify CHEMICAL and DISEASE entities from the question pairs. The identified CHEMICAL and DISEASE entities were further replaced by synonyms from the dictionary of concepts and synonyms created from Observational Medical Outcomes Partnership (OMOP) Common Data Model. OMOP has consolidated multiple vocabularies into a common format, and OMOP's Standardized Vocabularies contain all the code sets, terminologies, vocabularies, nomenclatures, lexicons, thesauri, ontologies, taxonomies, classifications, abstractions, and other

such data that are required. This saves researchers and developers from having to understand and handle multiple formats and conventions of the originating vocabularies. For back translation, we used Google Translate application programming interface to do back translation and Chinese as the middle language. We compared several middle languages and found that Chinese had the best performance in recognizing question similarity. The original source text and back-translated text were compared to find differences, if any, in which case the back-translated text was used as an augmented record. In these experiments, we used only BERT and Clinical BERT as our base language models because these 2 models were found to have the best performance on the original test data set.

The results of the experiments with the augmented training data are shown in [Table 5](#). We found that the data augmentation techniques did not improve the model's performance. Experiments with back translation–augmented data samples performed better than experiments with entity replacement–augmented data samples. By analyzing the augmented data samples, we found that although these 2 data augmentation techniques expanded the number of data samples, they introduced some noise samples to our training set. This could be explained by the complexity and specificity of the text in the regulatory domain.

Table 5. Model performance with augmented training data.

Model	Data split strategy	Accuracy (%)		
		Entity replacement	Back translation	Entity replacement+back translation
BERT ^a	1	79.33	77.33	79.33
BERT	2	79.33	85.33	77.33
Clinical BERT	1	88.66	86.66	82.66
Clinical BERT	2	84	89.33	88

^aBERT: Bidirectional Encoder Representations from Transformers.

Our study has some limitations. First, we only experimented with the BERT-based model in this study. Some other state-of-the-art pretrained language models, such as XLNet, T5, and GPT-2, also perform well in related NLP tasks. We will try other state-of-the-art models in our future studies. Second, we only had 150 pairs of questions in our test data set. If we had had a greater number of question pairs in our test data set, we would have better understood the performance of each model. Third, our manually labeled data set covers only a limited number and types of concepts in the regulatory domain. We should further our analysis by expanding the variety of question pairs.

Conclusions

This study used deep learning models to recognize question entailment in the pharmaceutical regulatory domain. As no previous study has used computational models to learn text in the regulatory domain, our study demonstrates the possibility of using state-of-the-art artificial intelligence–based NLP models to understand the regulatory text. We also attempted 2 data augmentation techniques, back translation and entity replacement, to increase the number of training samples. However, these 2 techniques did not improve the model's performance in this study.

Acknowledgments

The authors would like to thank Muthukumar Vaithianathan and Todd Sanger at Eli Lilly & Company. Muthu helped create the dictionary of entities and their synonyms from Observational Medical Outcomes Partnership Common Data Model. This dictionary is pivotal for creating the entity replacement–based augmented records discussed in this paper. Todd helped check and fix the grammar in the paper.

Data Availability

The Quora questions pairs data set [55] and clinical Recognizing Question Entailment data set [1] are publicly available. The regulatory Recognizing Question Entailment manually annotated data set generated during and analyzed during this study are not publicly available because of our company's data protection and confidentiality policy but are available from the corresponding author on reasonable request.

Conflicts of Interest

None declared.

References

1. Ben Abacha A, Demner-Fushman D. Recognizing Question Entailment for Medical Question Answering. AMIA Annu Symp Proc 2016;2016:310-318 [[FREE Full text](#)] [Medline: [28269825](#)]

2. dos Santos C, Barbosa L, Bogdanova D, Zadrozny B. Learning hybrid representations to retrieve semantically equivalent questions. In: Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing. 2015 Presented at: ACL-IJCNLP '15; July 26-31, 2015; Beijing, China p. 694-699 URL: <https://aclanthology.org/P15-2114.pdf> [doi: [10.3115/v1/p15-2114](https://doi.org/10.3115/v1/p15-2114)]
3. Nakov P, Hoogeveen D, Màrquez L, Moschitti A, Mubarak H, Baldwin T, et al. SemEval-2017 task 3: community question answering. In: Proceedings of the 11th International Workshop on Semantic Evaluation. 2017 Presented at: SemEval '17; August 3-4, 2017; Vancouver, Canada p. 27-48 URL: <https://aclanthology.org/S17-2003.pdf> [doi: [10.18653/v1/s17-2003](https://doi.org/10.18653/v1/s17-2003)]
4. Charlet D, Damnati G. SimBow at semeval-2017 task 3: soft-cosine semantic similarity between questions for community question answering. In: Proceedings of the 11th International Workshop on Semantic Evaluation. 2017 Presented at: SemEval '17; August 3-4, 2017; Vancouver, Canada URL: <https://aclanthology.org/S17-2051.pdf> [doi: [10.18653/v1/s17-2051](https://doi.org/10.18653/v1/s17-2051)]
5. Kunneman F, Ferreira TC, Kraemer E, van den Bosch A. Question similarity in community question answering: a systematic exploration of preprocessing methods and models. In: Proceedings of the 2019 International Conference on Recent Advances in Natural Language Processing. 2019 Presented at: RANLP '19; September 2-4, 2019; Varna, Bulgaria p. 593-601 URL: <https://aclanthology.org/R19-1070.pdf> [doi: [10.26615/978-954-452-056-4_070](https://doi.org/10.26615/978-954-452-056-4_070)]
6. Robertson S, Zaragoza H. The probabilistic relevance framework: BM25 and beyond. Found Trends Inf Retr 2009 Apr;3(4):333-389 [FREE Full text] [doi: [10.1561/1500000019](https://doi.org/10.1561/1500000019)]
7. Shah D, Lei T, Moschitti A, Romeo S, Nakov P. Adversarial domain adaptation for duplicate question detection. In: Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing. 2018 Presented at: EMNLP '18; October 31-November 4, 2018; Brussels, Belgium p. 1056-1063 URL: <https://aclanthology.org/D18-1131.pdf> [doi: [10.18653/v1/d18-1131](https://doi.org/10.18653/v1/d18-1131)]
8. Nguyen VT, Le AC, Nguyen HN. A model of convolutional neural network combined with external knowledge to measure the question similarity for community question answering systems. Int J Mach Learn 2021 May;11(3):194-201 [FREE Full text] [doi: [10.18178/ijmlc.2021.11.3.1035](https://doi.org/10.18178/ijmlc.2021.11.3.1035)]
9. Wang H, Zhu H, Wu H, Wang X, Han X, Xu T. A densely connected GRU neural network based on Coattention mechanism for Chinese rice-related question similarity matching. Agronomy 2021 Jun 27;11(7):1307 [FREE Full text] [doi: [10.3390/agronomy11071307](https://doi.org/10.3390/agronomy11071307)]
10. McCreery CH, Katariya N, Kannan A, Chablani M, Amatriain X. Effective transfer learning for identifying similar questions: matching user questions to COVID-19 FAQs. In: Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining. 2020 Presented at: KDD '20; July 6-10, 2020; Virtual Event p. 3458-3465 URL: <https://dl.acm.org/doi/10.1145/3394486.3412861> [doi: [10.1145/3394486.3412861](https://doi.org/10.1145/3394486.3412861)]
11. Ben Abacha A, Shivade C, Demner-Fushman D. Overview of the MEDIQA 2019 shared task on textual inference, question entailment and question answering. In: Proceedings of the 18th BioNLP Workshop and Shared Task. 2019 Presented at: BioNLP '19; August 1, 2019; Florence, Italy p. 370-379 URL: <https://aclanthology.org/W19-5039.pdf> [doi: [10.18653/v1/w19-5039](https://doi.org/10.18653/v1/w19-5039)]
12. Luo J, Zhang GQ, Wentz S, Cui L, Xu R. SimQ: real-time retrieval of similar consumer health questions. J Med Internet Res 2015 Feb 17;17(2):e43 [FREE Full text] [doi: [10.2196/jmir.3388](https://doi.org/10.2196/jmir.3388)] [Medline: [25689608](https://pubmed.ncbi.nlm.nih.gov/25689608/)]
13. Wang D, Nyberg E. CMU OAQA at TREC 2017 LiveQA: a neural dual entailment approach for question paraphrase identification. In: Proceedings of the 26th Text REtrieval Conference. 2017 Presented at: TREC '17; November 15-17, 2017; Gaithersburg, MD URL: <https://trec.nist.gov/pubs/trec26/papers/CMU-OAQA-QA.pdf>
14. Ben Abacha A, Demner-Fushman D. A question-entailment approach to question answering. BMC Bioinformatics 2019 Oct 22;20(1):511 [FREE Full text] [doi: [10.1186/s12859-019-3119-4](https://doi.org/10.1186/s12859-019-3119-4)] [Medline: [31640539](https://pubmed.ncbi.nlm.nih.gov/31640539/)]
15. Zhu W, Zhou X, Wang K, Luo X, Li X, Ni Y, et al. PANLP at MEDIQA 2019: pre-trained language models, transfer learning and knowledge distillation. In: Proceedings of the 18th BioNLP Workshop and Shared Task. 2019 Presented at: BioNLP '19; August 1, 2019; Florence, Italy p. 380-388 URL: <https://aclanthology.org/W19-5040.pdf> [doi: [10.18653/v1/w19-5040](https://doi.org/10.18653/v1/w19-5040)]
16. Devlin J, Chang MW, Lee K, Toutanova K. Bert: pre-training of deep bidirectional transformers for language understanding. In: Proceedings of the 17th Annual Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. 2019 Presented at: NAACL-HLT '19; June 2-7, 2019; Minneapolis, MN p. 4171-4186 URL: <https://aclanthology.org/N19-1423.pdf>
17. Liu X, He P, Chen W, Gao J. Multi-task deep neural networks for natural language understanding. In: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. 2019 Presented at: ACLN'19; July 28-August 2, 2019; Florence, Italy p. 4487-4496 URL: <https://aclanthology.org/P19-1441.pdf> [doi: [10.18653/v1/p19-1441](https://doi.org/10.18653/v1/p19-1441)]
18. Bhaskar SA, Rungta R, Route J, Nyberg E, Mitamura T. Sieg at MEDIQA 2019: multi-task neural ensemble for biomedical inference and entailment. In: Proceedings of the 18th BioNLP Workshop and Shared Task. 2019 Presented at: BioNLP '19; August 1, 2019; Florence, Italy p. 462-470 URL: <https://aclanthology.org/W19-5049.pdf> [doi: [10.18653/v1/w19-5049](https://doi.org/10.18653/v1/w19-5049)]
19. Sharma P, Roychowdhury S. IIT-KGP at MEDIQA 2019: recognizing question entailment using sci-Bert stacked with a gradient boosting classifier. In: Proceedings of the 18th BioNLP Workshop and Shared Task. 2019 Presented at: BioNLP '19; August 1, 2019; Florence, Italy p. 471-477 URL: <https://aclanthology.org/W19-5050.pdf> [doi: [10.18653/v1/w19-5050](https://doi.org/10.18653/v1/w19-5050)]

20. Pugaliya H, Saxena K, Garg S, Shalini S, Gupta P, Nyberg E, et al. Pentagon at MEDIQA 2019: multi-task learning for filtering and re-ranking answers using language inference and question entailment. arXiv. Preprint posted online July 1, 2019 [FREE Full text] [doi: [10.18653/v1/w19-5041](https://doi.org/10.18653/v1/w19-5041)]
21. Sarrouiti M, Ben Abacha A, Demner-Fushman D. Multi-task transfer learning with data augmentation for recognizing question entailment in the medical domain. In: Proceedings of the 9th International Conference on Healthcare Informatics. 2021 Presented at: ICHI '21; August 9-12, 2021; Victoria, BC p. 339-346 URL: <https://ieeexplore.ieee.org/document/9565717> [doi: [10.1109/ichi52183.2021.00058](https://doi.org/10.1109/ichi52183.2021.00058)]
22. Wang Y, Afzal N, Liu S, Rastegar-Mojarad M, Wang L, Shen F, et al. Overview of the BioCreative/OHNLN challenge 2018 task 2: clinical semantic textual similarity. In: Proceedings of the 2018 ACM International Conference on Bioinformatics, Computational Biology, and Health Informatics. 2018 Presented at: BCB '18; August 29-September 1, 2018; Washington, DC URL: https://www.researchgate.net/profile/Yanshan-Wang-2/publication/327424883_Overview_of_BioCreativeOHNLN_Challenge_2018_Task_2_Clinical_Semantic_Textual_Similarity/links/5b8eb6bda6fdcc1ddd0ebcce/Overview-of-BioCreative-OHNLN-Challenge-2018-Task-2-Clinical-Semantic-Textual-Similarity.pdf [doi: [10.1145/3233547.3233672](https://doi.org/10.1145/3233547.3233672)]
23. Lastra-Díaz JJ, García-Serrano A. A new family of information content models with an experimental survey on WordNet. Knowl Based Syst 2015 Nov;89:509-526 [FREE Full text] [doi: [10.1016/j.knosys.2015.08.019](https://doi.org/10.1016/j.knosys.2015.08.019)]
24. Hadj Taieb MA, Ben Aouicha M, Ben Hamadou A. Ontology-based approach for measuring semantic similarity. Eng Appl Artif Intell 2014 Nov;36:238-261 [FREE Full text] [doi: [10.1016/j.engappai.2014.07.015](https://doi.org/10.1016/j.engappai.2014.07.015)]
25. Elavarasi SA, Akilandeswari J, Menaga K. A survey on semantic similarity measure. Int J Res Advent Technol 2014;2(3):389-398.
26. Lastra-Díaz JJ, Goikoetxea J, Hadj Taieb MA, García-Serrano A, Ben Aouicha M, Agirre E. A reproducible survey on word embeddings and ontology-based methods for word similarity: linear combinations outperform the state of the art. Eng Appl Artif Intell 2019 Oct;85:645-665 [FREE Full text] [doi: [10.1016/j.engappai.2019.07.010](https://doi.org/10.1016/j.engappai.2019.07.010)]
27. Li Y, McLean D, Bandar ZA, O'Shea JD, Crockett K. Sentence similarity based on semantic nets and corpus statistics. IEEE Trans Knowl Data Eng 2006 Aug;18(8):1138-1150 [FREE Full text] [doi: [10.1109/tkde.2006.130](https://doi.org/10.1109/tkde.2006.130)]
28. Rubenstein H, Goodenough JB. Contextual correlates of synonymy. Commun ACM 1965 Oct;8(10):627-633 [FREE Full text] [doi: [10.1145/365628.365657](https://doi.org/10.1145/365628.365657)]
29. Marelli M, Menini S, Baroni M, Bentivogli L, Bernardi R, Zamparelli R. A SICK cure for the evaluation of compositional distributional semantic models. In: Proceedings of the 9th International Conference on Language Resources and Evaluation. 2014 Presented at: LREC '14; May 18-21, 2014; Reykjavik, Iceland p. 216-233 URL: http://www.lrec-conf.org/proceedings/lrec2014/pdf/363_Paper.pdf [doi: [10.3115/v1/s14-2001](https://doi.org/10.3115/v1/s14-2001)]
30. Agirre E, Banea C, Cardie C, Cer D, Diab M, Gonzalez-Agirre A, et al. SemEval-2015 task 2: semantic textual similarity, English, Spanish and pilot on interpretability. In: Proceedings of the 9th International Workshop on Semantic Evaluation. 2015 Presented at: SemEval '15; June 4-5, 2015; Denver, Colorado p. 252-263 URL: <https://aclanthology.org/S15-2045.pdf> [doi: [10.18653/v1/s15-2045](https://doi.org/10.18653/v1/s15-2045)]
31. Agirre E, Banea C, Cer D, Diab M, Gonzalez-Agirre A, Mihalcea R, et al. SemEval-2016 task 1: semantic textual similarity, monolingual and cross-lingual evaluation. In: Proceedings of the 10th International Workshop on Semantic Evaluation. 2016 Presented at: SemEval '16; June 16-17, 2016; San Diego, CA p. 497-511 URL: <https://aclanthology.org/S16-1081.pdf> [doi: [10.18653/v1/s16-1081](https://doi.org/10.18653/v1/s16-1081)]
32. Agirre E, Diab MT, Cer D, Gonzalez-Agirre A. SemEval-2012 task 6: a pilot on semantic textual similarity. In: Proceedings of the First Joint Conference on Lexical and Computational Semantics – Volume 1: Proceedings of the main conference and the shared task, and Volume 2: Proceedings of the 6th International Workshop on Semantic Evaluation. 2012 Presented at: SemEval '12; June 7-8, 2012; Montréal, QC p. 385-393 URL: <https://aclanthology.org/S12-1051.pdf> [doi: [10.18653/v1/s16-1081](https://doi.org/10.18653/v1/s16-1081)]
33. Agirre E, Cer D, Diab M, Gonzalez-Agirre A, Guo W. *SEM 2013 shared task: semantic textual similarity. In: Proceedings of the 2nd Joint Conference on Lexical and Computational Semantics (*SEM), Volume 1: Proceedings of the Main Conference and the Shared Task: Semantic Textual Similarity. 2013 Presented at: *SEM '13; June 13-14, 2013; Atlanta, GA p. 32-43 URL: <https://aclanthology.org/S13-1004.pdf> [doi: [10.3115/v1/s14-2010](https://doi.org/10.3115/v1/s14-2010)]
34. Cer D, Diab M, Agirre E, Lopez-Gazpio I, Specia L. SemEval-2017 task 1: semantic textual similarity - multilingual and cross-lingual focused evaluation. arXiv. Preprint posted online July 31, 2017 [FREE Full text] [doi: [10.18653/v1/s17-2001](https://doi.org/10.18653/v1/s17-2001)]
35. Kim S, Fiorini N, Wilbur WJ, Lu Z. Bridging the gap: incorporating a semantic similarity measure for effectively mapping PubMed queries to documents. J Biomed Inform 2017 Nov;75:122-127 [FREE Full text] [doi: [10.1016/j.jbi.2017.09.014](https://doi.org/10.1016/j.jbi.2017.09.014)] [Medline: [28986328](https://pubmed.ncbi.nlm.nih.gov/28986328/)]
36. Mohamed M, Oussalah M. SRL-ESA-TextSum: a text summarization approach based on semantic role labeling and explicit semantic analysis. Inf Process Manag 2019 Jul;56(4):1356-1372 [FREE Full text] [doi: [10.1016/j.ipm.2019.04.003](https://doi.org/10.1016/j.ipm.2019.04.003)]
37. Rakhlin A. Convolutional neural networks for sentence classification. GitHub. 2016. URL: <https://github.com/alexander-rakhlin/CNN-for-Sentence-Classification-in-Keras> [accessed 2021-12-11]
38. Janda HK, Pawar A, Du S, Mago V. Syntactic, semantic and sentiment analysis: the joint effect on automated essay evaluation. IEEE Access 2019;7:108486-108503 [FREE Full text] [doi: [10.1109/access.2019.2933354](https://doi.org/10.1109/access.2019.2933354)]

39. Zou WY, Socher R, Cer D, Manning CD. Bilingual word embeddings for phrase-based machine translation. In: Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing. 2013 Presented at: EMNLP '13; October 18-21, 2013; Seattle, WA p. 1393-1398 URL: <https://aclanthology.org/D13-1141.pdf>
40. Bordes A, Chopra S, Weston J. Question answering with subgraph embeddings. arXiv. Preprint posted online June 14, 2014 [FREE Full text] [doi: [10.3115/v1/d14-1067](https://doi.org/10.3115/v1/d14-1067)]
41. Lopez-Gazpio I, Maritxalar M, Gonzalez-Agirre A, Rigau G, Uria L, Agirre E. Interpretable semantic textual similarity: finding and explaining differences between sentences. Knowl Based Syst 2017 Mar 01;119:186-199 [FREE Full text] [doi: [10.1016/j.knosys.2016.12.013](https://doi.org/10.1016/j.knosys.2016.12.013)]
42. Amir S, Tanasescu A, Zighed DA. Sentence similarity based on semantic kernels for intelligent text retrieval. J Intell Inf Syst 2016 Nov 28;48(3):675-689 [FREE Full text] [doi: [10.1007/s10844-016-0434-3](https://doi.org/10.1007/s10844-016-0434-3)]
43. Benedetti F, Beneventano D, Bergamaschi S, Simonini G. Computing inter-document similarity with Context Semantic Analysis. Inf Syst 2019 Feb;80:136-147 [FREE Full text] [doi: [10.1016/j.is.2018.02.009](https://doi.org/10.1016/j.is.2018.02.009)]
44. Yang Y, Yuan S, Cer D, Kong SY, Constant N, Pilar P, et al. Learning semantic textual similarity from conversations. arXiv. Preprint posted online April 20, 2018 [FREE Full text] [doi: [10.18653/v1/w18-3022](https://doi.org/10.18653/v1/w18-3022)]
45. Wang Y, Liu F, Verspoor K, Baldwin T. Evaluating the utility of model configurations and data augmentation on clinical semantic textual similarity. In: Proceedings of the 19th SIGBioMed Workshop on Biomedical Language Processing. 2020 Presented at: BioNLP '20; July 9, 2020; Virtual Event p. 105-111 URL: <https://aclanthology.org/2020.bionlp-1.11.pdf> [doi: [10.18653/v1/2020.bionlp-1.11](https://doi.org/10.18653/v1/2020.bionlp-1.11)]
46. Xiong Y, Chen S, Qin H, Cao H, Shen Y, Wang X, et al. Distributed representation and one-hot representation fusion with gated network for clinical semantic textual similarity. BMC Med Inform Decis Mak 2020 Apr 30;20(Suppl 1):72 [FREE Full text] [doi: [10.1186/s12911-020-1045-z](https://doi.org/10.1186/s12911-020-1045-z)] [Medline: [32349764](https://pubmed.ncbi.nlm.nih.gov/32349764/)]
47. Yang X, He X, Zhang H, Ma Y, Bian J, Wu Y. Measurement of semantic textual similarity in clinical texts: comparison of transformer-based models. JMIR Med Inform 2020 Nov 23;8(11):e19735 [FREE Full text] [doi: [10.2196/19735](https://doi.org/10.2196/19735)] [Medline: [33226350](https://pubmed.ncbi.nlm.nih.gov/33226350/)]
48. Sakata W, Shibata T, Tanaka R, Kurohashi S. FAQ retrieval using query-question similarity and BERT-based query-answer relevance. In: Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval. 2019 Presented at: SIGIR '19; July 21-25, 2019; Paris, France p. 1113-1116 URL: <https://dl.acm.org/doi/10.1145/3331184.3331326> [doi: [10.1145/3331184.3331326](https://doi.org/10.1145/3331184.3331326)]
49. Minaee S, Liu Z. Automatic question-answering using a deep similarity neural network. In: Proceedings of the 2017 IEEE Global Conference on Signal and Information Processing. 2017 Presented at: GlobalSIP '17; November 14-16, 2017; Montreal, QC p. 923-927. [doi: [10.1109/globalsip.2017.8309095](https://doi.org/10.1109/globalsip.2017.8309095)]
50. Uva A, Bonadiman D, Moschitti A. Injecting relational structural representation in neural networks for question similarity. arXiv. Preprint posted online June 20, 2018 [FREE Full text] [doi: [10.18653/v1/p18-2046](https://doi.org/10.18653/v1/p18-2046)]
51. Cai Y, Zhang Q, Lu W, Che X. A hybrid approach for measuring semantic similarity based on IC-weighted path distance in WordNet. J Intell Inf Syst 2017 Sep 5;51(1):23-47 [FREE Full text] [doi: [10.1007/s10844-017-0479-y](https://doi.org/10.1007/s10844-017-0479-y)]
52. Alsentzer E, Murphy JR, Boag W, Weng WH, Jin D, Naumann T, et al. Publicly available clinical BERT embeddings. arXiv. Preprint posted online April 06, 2019 [FREE Full text] [doi: [10.1090/mbk/121/79](https://doi.org/10.1090/mbk/121/79)]
53. Lee J, Yoon W, Kim S, Kim D, Kim S, So CH, et al. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. Bioinformatics 2019 Sep 10;36(4):1234-1240 [FREE Full text] [doi: [10.1093/bioinformatics/btz682](https://doi.org/10.1093/bioinformatics/btz682)]
54. Peng Y, Yan S, Lu Z. Transfer learning in biomedical natural language processing: an evaluation of BERT and ELMo on ten benchmarking datasets. arXiv. Preprint posted online June 13, 2019 [FREE Full text] [doi: [10.18653/v1/w19-5006](https://doi.org/10.18653/v1/w19-5006)]
55. Iyer S, Dandekar N, Csernai K. First quora dataset release: question pairs. Quora. 2017. URL: <https://quoradata.quora.com/First-Quora-Dataset-Release-Question-Pairs> [accessed 2023-10-20]
56. Kumar A, Bhattamishra S, Bhandari M, Talukdar P. Submodular optimization-based diverse paraphrasing and its effectiveness in data augmentation. In: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. 2019 Presented at: NAACL '19; June 2-7, 2019; Minneapolis, Minnesota p. 3609-3619 URL: <https://aclanthology.org/N19-1363.pdf> [doi: [10.18653/v1/n19-1363](https://doi.org/10.18653/v1/n19-1363)]
57. Sennrich R, Haddow B, Birch A. Improving neural machine translation models with monolingual data. arXiv. Preprint posted online November 20, 2015 [FREE Full text] [doi: [10.18653/v1/p16-1009](https://doi.org/10.18653/v1/p16-1009)]
58. Yang Y, Malaviya C, Fernandez J, Swayamdipta S, Le Bras R, Wang JP, et al. Generative data augmentation for commonsense reasoning. arXiv. Preprint posted online April 24, 2020 [FREE Full text] [doi: [10.18653/v1/2020.findings-emnlp.90](https://doi.org/10.18653/v1/2020.findings-emnlp.90)]
59. Wei J, Zou K. EDA: easy data augmentation techniques for boosting performance on text classification tasks. arXiv. Preprint posted online January 31, 2019 [FREE Full text] [doi: [10.18653/v1/d19-1670](https://doi.org/10.18653/v1/d19-1670)]
60. Chen H, Ji Y, Evans D. Finding friends and flipping frenemies: automatic paraphrase dataset augmentation using graph theory. arXiv. Preprint posted online November 3, 2020 [FREE Full text] [doi: [10.18653/v1/2020.findings-emnlp.426](https://doi.org/10.18653/v1/2020.findings-emnlp.426)]
61. Kang T, Perotte A, Tang Y, Ta C, Weng C. UMLS-based data augmentation for natural language processing of clinical research literature. J Am Med Inform Assoc 2021 Mar 18;28(4):812-823 [FREE Full text] [doi: [10.1093/jamia/ocaa309](https://doi.org/10.1093/jamia/ocaa309)] [Medline: [33367705](https://pubmed.ncbi.nlm.nih.gov/33367705/)]

62. Zhang H, Cisse M, Dauphin YN, Lopez-Paz D. mixup: beyond empirical risk minimization. arXiv. Preprint posted online October 25, 2017 [[FREE Full text](#)]
63. Chen J, Yang Z, Yang D. MixText: linguistically-informed interpolation of hidden space for semi-supervised text classification. arXiv. Preprint posted online April 25, 2020 [[FREE Full text](#)] [doi: [10.18653/v1/2020.acl-main.194](https://doi.org/10.18653/v1/2020.acl-main.194)]
64. Radford A, Wu J, Child R, Luan D, Amodei D, Sutskever I. Language Models are Unsupervised Multitask Learners. OpenAI blog 2019;1(8):9 [[FREE Full text](#)]
65. Neumann M, King D, Beltagy I, Ammar W. ScispaCy: fast and robust models for biomedical natural language processing. arXiv. Preprint posted online February 20, 2019 [[FREE Full text](#)] [doi: [10.18653/v1/w19-5034](https://doi.org/10.18653/v1/w19-5034)]

Abbreviations

ABCNN: attention-based convolutional neural network
AI: artificial intelligence
AUC: area under the receiver operating characteristic curve
BERT: Bidirectional Encoder Representations from Transformers
Bi-LSTM: bidirectional long short-term memory
BLUE: Biomedical Language Understanding Evaluation
C-RQE: clinical Recognizing Question Entailment
DNN: deep neural network
FAQ: frequently asked questions
MAD: manually annotated data set
MIMIC-III: Medical Information Mart for Intensive Care
NLI: natural language interface
NLP: natural language processing
OMOP: Observational Medical Outcomes Partnership
QA: question answering
RA: regulatory affairs
RQE: Recognizing Question Entailment
STS: semantic textual similarity
TE: textual entailment

Edited by K El Emam; submitted 12.10.22; peer-reviewed by W Klement, S Sharma; comments to author 16.02.23; revised version received 29.03.23; accepted 02.07.23; published 26.09.23.

Please cite as:

Saraswat N, Li C, Jiang M

Identifying the Question Similarity of Regulatory Documents in the Pharmaceutical Industry by Using the Recognizing Question Entailment System: Evaluation Study

JMIR AI 2023;2:e43483

URL: <https://ai.jmir.org/2023/1/e43483>

doi: [10.2196/43483](https://doi.org/10.2196/43483)

PMID:

©Nidhi Saraswat, Chuqin Li, Min Jiang. Originally published in JMIR AI (<https://ai.jmir.org>), 26.09.2023. This is an open-access article distributed under the terms of the Creative Commons Attribution License (<https://creativecommons.org/licenses/by/4.0/>), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR AI, is properly cited. The complete bibliographic information, a link to the original publication on <https://www.ai.jmir.org/>, as well as this copyright and license information must be included.

Original Paper

Physicians' and Machine Learning Researchers' Perspectives on Ethical Issues in the Early Development of Clinical Machine Learning Tools: Qualitative Interview Study

Jane Paik Kim¹, PhD; Katie Ryan¹, BA, MA; Max Kasun¹, BA; Justin Hogg¹, BA; Laura B Dunn², MD; Laura Weiss Roberts¹, MA, MD

¹Department of Psychiatry and Behavioral Sciences, Stanford University School of Medicine, Palo Alto, CA, United States

²Department of Psychiatry, University of Arkansas for Medical Sciences, Arkansas, CA, United States

Corresponding Author:

Jane Paik Kim, PhD

Department of Psychiatry and Behavioral Sciences

Stanford University School of Medicine

1070 Arastradero Road

Palo Alto, CA, 94304-1125

United States

Phone: 1 6507368996

Email: jane.pkim@stanford.edu

Abstract

Background: Innovative tools leveraging artificial intelligence (AI) and machine learning (ML) are rapidly being developed for medicine, with new applications emerging in prediction, diagnosis, and treatment across a range of illnesses, patient populations, and clinical procedures. One barrier for successful innovation is the scarcity of research in the current literature seeking and analyzing the views of AI or ML researchers and physicians to support ethical guidance.

Objective: This study aims to describe, using a qualitative approach, the landscape of ethical issues that AI or ML researchers and physicians with professional exposure to AI or ML tools observe or anticipate in the development and use of AI and ML in medicine.

Methods: Semistructured interviews were used to facilitate in-depth, open-ended discussion, and a purposeful sampling technique was used to identify and recruit participants. We conducted 21 semistructured interviews with a purposeful sample of AI and ML researchers (n=10) and physicians (n=11). We asked interviewees about their views regarding ethical considerations related to the adoption of AI and ML in medicine. Interviews were transcribed and deidentified by members of our research team. Data analysis was guided by the principles of qualitative content analysis. This approach, in which transcribed data is broken down into descriptive units that are named and sorted based on their content, allows for the inductive emergence of codes directly from the data set.

Results: Notably, both researchers and physicians articulated concerns regarding how AI and ML innovations are shaped in their early development (ie, the *problem formulation* stage). Considerations encompassed the assessment of research priorities and motivations, clarity and centeredness of clinical needs, professional and demographic diversity of research teams, and interdisciplinary knowledge generation and collaboration. Phase-1 ethical issues identified by interviewees were notably interdisciplinary in nature and invited questions regarding how to align priorities and values across disciplines and ensure clinical value throughout the development and implementation of medical AI and ML. Relatedly, interviewees suggested interdisciplinary solutions to these issues, for example, more resources to support knowledge generation and collaboration between developers and physicians, engagement with a broader range of stakeholders, and efforts to increase diversity in research broadly and within individual teams.

Conclusions: These qualitative findings help elucidate several ethical challenges anticipated or encountered in AI and ML for health care. Our study is unique in that its use of open-ended questions allowed interviewees to explore their sentiments and perspectives without overreliance on implicit assumptions about what AI and ML currently are or are not. This analysis, however, does not include the perspectives of other relevant stakeholder groups, such as patients, ethicists, industry researchers or representatives, or other health care professionals beyond physicians. Additional qualitative and quantitative research is needed to reproduce and build on these findings.

KEYWORDS

artificial intelligence; machine learning; ethical considerations; qualitative study; qualitative; ethic; ethics; ethical; perspective

Introduction

Background

Innovation in the field of machine learning (ML) within artificial intelligence (AI) is accelerating in medicine, with more US Food and Drug Administration (FDA) approvals for algorithms and devices in 2022 than in any prior year [1,2]. As algorithm research, tool development, and clinical implementation proceed, AI and ML innovations stand to benefit many domains of medicine, from enhanced classification systems for clinical diseases and syndromes to highly individualized patient care, encompassing prediction, diagnosis, and treatment [3,4]. In parallel, ethics governance has been recognized as a priority and standard for the advancement of AI and ML in medicine, with recent guidance emerging from working groups, expert meetings, and scholarly work [5-8]. There is now a wide agreement that a failure to anticipate ethical issues threatens to compromise public trust in medicine and, ultimately, its embrace of AI and ML and their promise to improve human health [9].

Attention to ethical challenges in medical AI and ML has increased sharply in recent years in response to evidence showing that clinical AI and ML tools may offer limited generalizability and reproducibility [10,11], low rates of successful clinical adoption [12], and algorithmic bias [13]. Although the accompanying risks may not always be immediately obvious, past examples teach us that premature clinical integration of innovative tools can lead to *runaway diffusion* of risks to patients in clinical research and routine care, ranging from reduced benefit of the tools to outright harms, which rapidly become harder to address as tools become more widespread and ingrained in clinical processes [8]. Critical and timely work in clinical ethics has emerged to proactively meet emerging challenges through the articulation of possible frameworks and recommendations and with the benefit of instructive early case studies [6,14-17]. McCradden et al [6], for example, proposed an oversight procedure for medical AI to help bridge the *AI chasm* created by the divergent ethical and methodological norms of the clinical and computer sciences.

In addition to this foundational work, there is agreement that predicting and meeting new ethical challenges will require seeking, analyzing, and incorporating the perspectives of professionals who work along the full pipeline of AI and ML innovation (ie, key stakeholders) [18-20]. Seeking the perspectives of stakeholders and generating knowledge based on their perspectives serves to test the veracity of the assumptions made about their views, identify human factors that could present barriers to implementation, and better understand clinical needs. Bringing awareness to the areas of translation where developers' intentions may not align with the goals of end users may serve to minimize ethical "strain," as noted by Char et al [21]. This work is critical for forming comprehensive ethical guidance and responding constructively to differing normative views on AI and ML innovation.

Early stakeholder research regarding medical AI and ML primarily sought physicians' and patients' views regarding AI and ML tool implementation in clinical practice. It has yielded insights into concerns such as generalizability, algorithmic fairness, and clinical fit, as well as a range of ethical concerns that remain unclarified or unaddressed [22-25]. For instance, clinicians have reported uncertainty about their ability to collaborate effectively with AI and ML tools in clinics, given the numerous time and resource constraints of clinical ecosystems [25]. Patients expressed reservations about consenting to share health data for AI and ML research purposes and resistance to prognostic AI and ML systems that determine treatment admission without provider-patient dialogue [22]. Other reported considerations include end-users' perceptions of algorithms' utility, the potential for overreliance on algorithms when performing clinical tasks, users' lack of knowledge of the rules governing algorithms, and disruptions of existing clinical infrastructure, workflows, and configurations of care teams [15,16,23,26].

A major gap in the current stakeholder literature on ethics in AI and ML is that it has not frequently sought the perspectives of other key stakeholders such as AI and ML researchers and developers [27]. Furthermore, because most of this work has so far focused on perspectives regarding the implementation and use of specific clinical tools, few studies have analyzed stakeholder views on the ethical challenges that they perceive in other phases of the innovation pipeline, including conceptualization and development [28,29]. One noteworthy exception is an interview study of 19 informatics leaders at US academic medical centers, which found that leaders perceived efforts to build interdisciplinary consensus and define clinical needs as necessary before the clinical implementation phase [30]. Although input seeking from end users to inform upstream development has been conceived as potentially helpful in closing the implementation gap, stakeholder research has been underused as an empirical method for developing comprehensive ethics guidance [31].

Objectives

Therefore, the purpose of this study was to describe the landscape of ethical issues that AI and ML researchers and physicians with professional exposure to AI or ML tools observe or anticipate in the development and use of AI or ML in medicine. This report is the first in a series of papers to describe findings from a larger study in which we conducted open-ended, in-depth interviews with multiple stakeholder groups, including AI and ML researchers, physicians, ethicists, and patients. In this study, we focus on the perspectives of AI and ML researchers and physicians with professional exposure to AI or ML. Through an open-ended discussion, we aimed to identify ethical considerations that may not be as frequently elevated in the literature, potentially because of a hyperfocus on already known issues. Given the lack of prior work involving these 2 stakeholder groups, we had no a priori hypotheses about

common or divergent perspectives. Rather, we sought to describe the current landscape of ethical considerations.

Methods

Study Design

The purpose of this study, which is part of a broader project (National Center for Advancing Translational Sciences R01-TR-003505) studying the influence of AI and ML tools on clinical decision-making, was to describe the views of AI and ML researchers and physicians regarding ethical considerations they have encountered or anticipated in the development, refinement, and application of AI and ML in medicine [32]. A qualitative descriptive approach was applied in the design and completion of this study, as this method aims to describe specific experiences or perceptions using language directly from the data and is well suited for topics that have been minimally studied previously [33,34].

Semistructured interviews, which are a common method of data collection in qualitative research, were used in this study to facilitate in-depth, open-ended discussions [35]. The interview questions were intentionally broad in scope to allow participants the opportunity to address the topics that they personally found the most significant, as opposed to responding to topics defined a priori by our research team. As participants were not necessarily trained in ethics or familiar with the associated vocabulary, questions regarding the benefits, risks, and unintended consequences of AI and ML were included to encourage them to consider broader challenges related to their work that could potentially have ethical implications. Interview guides were tailored for each participant group (eg, physician and researcher), keeping in mind their professional backgrounds and contextual information. Ultimately, interview guides did not vary drastically from group to group (Textbox 1). After the first few interviews, the questions were slightly revised for clarity, based on feedback. Relevant follow-up questions were asked in response to participants' replies to the primary questions.

Textbox 1. Open-ended questions asked in interviews.

Asked of researchers

1. How would you describe your work? Can you give specific examples of recent work? What is the value of your work in the field?
2. What are some of your personal observations and experiences regarding the use of machine learning (ML) in medicine? Are there any special ethical issues you have encountered in the development of algorithms for medicine?
3. Do you have an example from your day-to-day work of algorithmic development that may have ethical implications?
4. Can you think of any unintended consequences in the application of ML algorithms in medicine?
5. Are there any other areas in the field of computer science that you work in that we have not covered yet in our conversation? Are there different ethical issues in this subfield compared with ML?
6. Do you believe that there are limits to what ML can accomplish in medicine?
7. What are your aspirations (or predictions) for your field? Do you anticipate any ethical issues?

Asked of physicians

1. How would you describe your work? Can you give me an example of what your average day looks like, or describe a few of the recent projects that you have been working on?
2. Can you describe any first-hand experiences that you have had using machine ML or artificial intelligence (AI) applications within health care?
3. What are your impressions or observations about the use of ML or AI applications in health care? Are there any special ethical issues that you have encountered or considered when it comes to using ML or AI applications in health care?
4. What do you think are some of the potential benefits of using ML or AI applications in health care? What do you think may be some of the unintended consequences?
5. What are your hopes or aspirations when it comes to ML or AI applications in health care? Do you anticipate any ethical issues?
6. How do you think the use of ML or AI applications will impact the jobs of doctors? Do you think it will have any impact on the patient-provider relationship?
7. What recommendations do you have for developers who are interested in creating ML or AI applications for the field of medicine?

Participants and Procedures

A purposeful sampling technique was used to identify and recruit the participants. Purposeful sampling is common in qualitative description research and involves identifying and recruiting specific individuals who are especially knowledgeable about the topic being studied [36]. For this project, we sought to interview researchers who had experience in developing AI or ML tools for use in medicine, and physicians who had

experience developing or using such tools. By consulting the relevant literature and seeking recommendations from experts in the fields of AI, ML, medicine, and AI ethics, we identified 61 candidates (33 researchers and 28 physicians) from 10 US academic institutions that met these criteria.

Recruitment e-mails containing details about our project and an electronic interest form were sent to these 61 potential participants. A total of 29 potential participants submitted an

electronic interest form. Of these, 21 (10 researchers and 11 physicians) scheduled and completed an interview. Interviews continued until content saturation was reached, that is, when additional data did not lead to the emergence of new or original ideas or themes [37-39]. The final cohort of participants was affiliated with 6 different US academic institutions and represented a variety of academic departments, including medicine, biomedical informatics, engineering, computer science, radiology, psychiatry, and surgery. All participants in the researcher group held master's degrees or higher in computer science or a related field, and all participants in the physician group held MDs. The complete demographic information of the participants is available in [Table 1](#).

Web-based interviews were conducted between November 2020 and April 2021 using Zoom (Zoom Video Communications). A PDF copy of the institutional review board–approved informed consent form was sent to all potential participants before their scheduled interview date. On the day of the interview, the interviewers verbally reviewed the content of the informed consent form with potential participants and answered any questions before obtaining verbal consent and beginning the interview. Interviews were conducted by 1 of our team's 4 trained interviewers and lasted 52 minutes, 6 seconds on average, ranging from 29 to 95 minutes (SD 15 min 54 s). The interviews were audio recorded. The participants were compensated in the form of an electronic gift card to appreciate their time and effort.

Table 1. Study population characteristics by the participant group^a.

Characteristics	Researcher (n=10)	Physician (n=11)	Overall (N=21)
Gender, n (%)			
Men	4 (40)	8 (73)	12 (57)
Women	6 (60)	3 (27)	9 (43)
Age (y)			
Value, mean (SD)	31.6 (3.91)	48.6 (17.7)	41.0 (15.7)
Value, median (IQR)	31.0 (27.0-37.0)	44.0 (30.0-93.0)	35.5 (27.0-93.0)
Race, n (%)			
African American or Black	1 (10)	1 (9)	2 (10)
Asian	4 (40)	5 (45)	9 (43)
White	4 (40)	3 (27)	7 (33)
Other	1 (10)	2 (18)	3 (14)
Ethnicity, n (%)			
Not Hispanic or Latino	10 (100)	9 (82)	19 (90)
Hispanic or Latino	0 (0)	2 (18)	2 (10)
Degree, n (%)			
Doctor of Medicine	0 (0)	5 (45)	5 (24)
Doctor of Medicine or Doctor of Philosophy	0 (0)	5 (45)	5 (24)
Doctor of Philosophy or equivalent	5 (50)	1 (9)	6 (29)
Master's	5 (50)	0 (0)	5 (24)

^aNote: 1 participant did not report age. Ten physician participants were Doctors of Medicine; 1 was a Doctor of Philosophy clinical psychologist.

Data Coding and Analysis

The interviews were transcribed and deidentified by the members of our research team. Data analysis was guided by the principles of qualitative content analysis [40]. This approach, in which transcribed data are broken down into descriptive units, which are named and sorted based on their content, allows for the inductive emergence of codes directly from the data set [41]. After the transcription of the interviews, open coding was performed for each transcript by 2 authors. The authors independently highlighted the substantive interview content and suggested descriptive codes for this content. The authors then met as a group to review and discuss these preliminary codes and refine their names and definitions. All transcripts were then

rereviewed by the 2 authors and coded using preliminary codes. The authors compared the coded units, refined the code names and definitions, and drafted the final version of the codebook, which contained 30 descriptive codes derived directly from the content of the interviews.

The transcripts and codebook were uploaded to NVivo 1.0 (QSR International) for final coding, which was completed by a single author (KR) and reviewed by the principal investigator (JPK). The full team contributed to the analysis, which involved assessing the coded units and developing categories and themes that described the coded content.

Ethics Approval

This study obtained approval from the Stanford University Institutional Review Board before the start of research (approved protocol # 58118).

Results

Overview

Qualitative content analysis was performed on the full data set, resulting in the identification of 30 inductive codes that described participants' considerations relating to 3 distinct phases of AI and ML development for medicine: the problem formulation phase (phase 1), the algorithm development phase (phase 2), and the clinical implementation phase (phase 3; Figure 1).

Notably, 18 (86%) out of 21 researcher and physician interviewees addressed considerations related to phase 1. We describe this phase as the *problem formulation* phase, but it has been denoted in other literature as the *topic selection*, *need identification*, or *project definition* phase. This phase involves processes such as identifying health care needs that could be amenable to AI or ML solutions and formulating the scientific questions relevant to solving those needs.

Of the 30 inductive codes, 7 (23%) were primarily affiliated with phase 1; from these 7 codes, 5 major themes emerged (Figure 2). Within these themes, which are described in detail in this paper, interviewees identified a set of tightly interrelated phase-1 considerations that they perceived as having influence on the ethical dimensions of AI and ML research in medicine. Inductive codes and themes relating to phases 2 and 3 were also identified; due to the scope of the current paper, the analysis of these findings will be presented in a subsequent report.

Figure 1. Phases of medical artificial intelligence and machine learning development as described by participants, and related inductive codes.

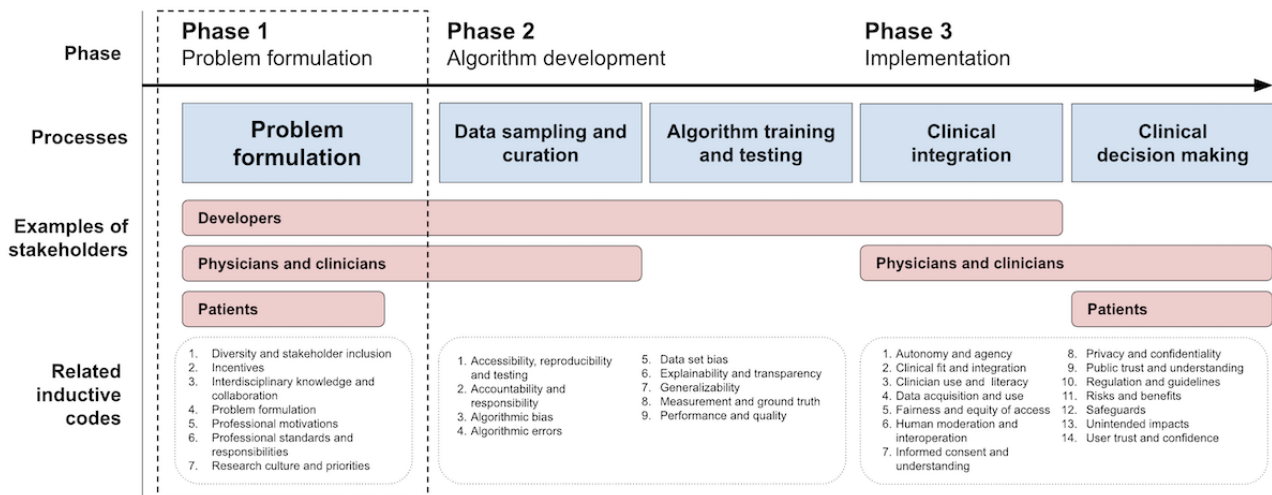


Figure 2. Inductive codes and themes describing researcher and physician views on phase 1 ethical considerations in medical artificial intelligence (AI) and machine learning (ML). For purposes of clarity, the codes “Professional motivations” and “Professional standards and responsibilities” were combined into a single row (“Professional motivations and standards”) for this figure.

Inductive codes	Number of interviews referenced in			Themes developed from inductive codes	Codes used in theme development
	Researcher (N=10), n (%)	Physician (N=11), n (%)	Total (N=21), n (%)		
Problem formulation ●	7 (70)	7 (64)	14 (66)	Assessing priorities and motivations in the development of AI and ML for medicine	● □ □ ◆ ⊕ ♥
Interdisciplinary knowledge and collaboration ▲	6 (60)	5 (45)	11 (52)	Evaluating the need for AI and ML	● □ ■ □ ⊕ □
Diversity and stakeholder inclusion ■	4 (40)	5 (45)	9 (43)	Developing AI and ML tools that have clinical value	● ▲ ■ ◆ ⊕ □
Research culture and priorities ◆	6 (60)	3 (27)	9 (43)	Engaging diverse stakeholder perspectives	● ▲ ■ □ □ □
Incentives ⊕	2 (20)	5 (45)	7 (33)	Advancing interdisciplinary literacy and ethical alignment	● ▲ □ ◆ □ ♥
Professional motivations and standards ♥	2 (20)	5 (45)	7 (33)		

Assessing Priorities and Motivations in the Development of AI and ML for Medicine

When asked to discuss the ethical tensions they experienced or expected in medical AI or ML, both researchers and physicians described the extent to which different priorities between AI and ML and medicine could introduce ethical tensions as the 2 fields increasingly interact. As one researcher summarized, “the Silicon Valley ‘move fast, break things’ mantra doesn’t really work in healthcare” (participant 18, researcher). They drew clear contrasts between the priorities of AI and ML (ie, rapid innovation and development) and those of medicine (ie, reducing suffering and hardship associated with health disorders and conditions) and felt that value misalignment was especially likely to undermine innovation in the context of health care:

You’re dealing with a new technology, and you’re kind of straddling between innovation and patient care, and often those do not align with each other. You’re trying to, obviously, innovate to improve patient care, but at the same time, the innovation part of it may not necessarily be the best for the patient, or may not necessarily be the thing that is most needed at the moment. [Participant 15, physician]

Both physicians and researchers reflected on the nature of their research communities’ interests in AI and ML innovation, with some suggesting that they may sometimes be unduly influenced by factors unrelated to obligations to patient care. As one physician emphasized:

Everyone wants to think that their innovation is going to be the one that actually changes health care, but ultimately you have to be mindful of, “Am I doing this because I want to innovate or am I doing this because I really want to prove this one process or take care of the patient?” [Participant 15, physician]

Another researcher reflected on how the pursuit of innovation for its own sake can lead to AI and ML solutions that may yield technical and intellectual insights and a sense of accomplishment but may not always be grounded in or aligned with clinical needs:

The “so what?” also becomes part of the problem. I think a lot of machine learning people, myself included, have a tendency to think “Oh this is a great machine learning problem because...it is a really cool intellectual problem.” But I think it becomes a human problem once you think about the fact that this model could be deployed in the hospital. [Participant 10, researcher]

Several interviewees expressed wariness about the ease with which researchers and physicians may be motivated by their beliefs and attitudes regarding the promise and potential of AI and ML in medicine. One physician commented on how “often, with technologies like AI that garner a lot of attention and funding, there is a tendency to be driven by this desire to use the technology just because it’s a technology that’s interesting” (participant 15, physician). Another, looking back, saw their optimism decline somewhat over their research career:

Before I started doing any machine learning research I felt like machine learning was this sort of holy grail that was going to solve every research question. [Participant 22, physician]

Evaluating the Need for AI and ML

Reflecting on the allure of AI and ML as an innovative field, interviewees expressed concern regarding the potential overproliferation of AI and ML tools in medicine for needs that could be better addressed with other technologies or interventions. Interviewees in both groups elevated the importance of performing an early assessment of the need for AI and ML and the value that it may add to specific medical contexts, that is, “Whether this is something that needs machine learning in the first place” (participant 22, physician)—as well considering, “At what cost is the question?” (participant 06, researcher) and whether a given AI or ML tool is “the right tool for the job” (participant 20, physician). One interviewee highlighted the importance of thinking about AI as one of many other innovative fields that can broadly apply to medicine in the interests of patients and humanity:

Think of AI as just an enabling technology like anything else. We don’t do anything for the sake of using the electronic health record. The electronic health record is a tool that allows us to take care of a patient, so AI should be the same thing. You should never do something for the sake of using AI. It should always be that we’re trying to solve this problem, we can’t solve it using existing tools so let’s see if AI, if prediction, could allow us to formulate a better solution...You always have to ground yourself and then go back to, “this is just one of many technologies that I use, but ultimately I have to focus on solving the problem of taking care of the patient in front of me.” [Participant 15, physician]

Developing AI and ML Tools That Have Clinical Value

Beyond initial assessments of the general need for AI and ML solutions, interviewees in both groups emphasized that it was just as important to evaluate the potential clinical value of the AI and ML tools under investigation. Several researchers expressed concerns about the proliferation of AI and ML tools that are not sufficiently evidence-based, that is, those created by developers who “[say] that [they]’re going to save time in the clinic, and then that’s not possible—[They] have no evidence to show that that’s the case” (participant 09, researcher). A recommendation to promote more robust AI and ML development proposed by members of both groups, in the words of a physician, was to “ground ourselves by being problem-focused” (participant 15, physician). In particular, they cautioned against presuming the benefits of AI and ML in a given problem in medicine, which they felt could reduce the likelihood of creating tools with clinical value:

We can’t just focus on building these new tools, but we need to think about the context in which they’re going to be used. We can’t think only about overall metrics...We really need to be prioritizing work that is actually meaningful, grounded in real problems as

far as how machine learning has been used in healthcare. {Participant 33, researcher}

Interviewees in both groups referenced institutional and structural factors in research and academia that they felt could promote the development of tools that might not ultimately prove useful in clinical settings. Specifically, they discussed the impact of academic competition in the medical and AI and ML research communities, which is important for advancing a career but may not adequately prioritize the development of practical tools. Researchers and physicians agreed that “There is a disconnect because the traditional ways of academic advancement—publications, grants—reward publication of algorithms and capabilities, and papers. There are many, many, many algorithms, but we don’t see many translate into clinical practice” (participant 23, physician). They stressed their desire “to see the incentives in academia and elsewhere change so that people can really invest in solving real problems instead of just churning out these publishable units.” (participant 33, researcher).

Several researchers felt that the rapid advancement of AI and ML in medicine could paradoxically slow progress, with several mentioning that stepwise approaches to development should be prioritized (“Sometimes...we do not currently have the ability to arrive at a satisfactory solution...Sometimes we are aiming for [step] three or four and we should just start with one” [participant 13, researcher]). They posited that more fundamental AI and ML research may yield greater benefits in the field’s current state (“I tend to believe that the most “boring” work is the work that pays off the most” [participant 18, researcher]). Likewise, some observed that more fundamental work is often overlooked in favor of research topics which follow “the new trend” (participant 10, researcher). They expressed frustration that the AI and ML research community appears to prioritize and reward “trend-hopping” (participant 33, researcher), whereas the work needed to create benefit for clinical populations may be neglected or underfunded (“we’re probably not putting our energy in the right place” [participant 10, researcher]).

Engaging Diverse Stakeholder Perspectives

Interviewees from both groups commented on the ways in which the composition of the research team—in terms of professional role (ie, developers, hospital administrators, physicians, and patients) and demographic characteristics (ie, race, gender, and ethnicity)—can impact the clinical utility of AI and ML solutions downstream:

Developers - definitely [work] with clinicians and communities and patients from the start. Because it's tempting, not just as a developer but also as a researcher, to feel like 'I have this really cool idea and I have this really cool algorithm and I'm just going to build it and then test it.' But that's a little bit of a disaster or a little bit of a risk of missing a lot of issues, or solving a problem that doesn't have to be solved, or not solving the right problem. {Participant 22, physician}

Interviewees in both groups emphasized the need to involve diverse collaborators “from the very beginning if we want to actually build something practically useful” (participant 13,

researcher), with a special emphasis on including individuals who may have “certain lived experience, [who] are going to be able to identify some things that others wouldn’t see” (participant 14, researcher).

Multiple interviewees similarly addressed how the background and demographic composition of a research team can impact the types of AI and ML projects that are advanced in medicine, with one researcher commenting “I think a lot of it, for better or for worse, is motivated by personal excitement and motivation, and not direct thinking about what kind of problems we should be addressing. That reason to me is the most concrete motivation for why we should have diversity in the field” (participant 10, researcher). Other researchers emphasized the importance of considering “Who’s asking the questions?” (participant 14, researcher), as well as, including those “who haven’t had the privilege to ask the questions, who haven’t been empowered to be able to ask the questions” (participant 33, researcher). Some expressed concerns that a lack of diversity among investigators and research teams could skew research directions and minimize the concerns of underrepresented and marginalized groups:

We see a lot of that in machine learning: It's not driven by what's a real question in the communities...It's driven by an idea that somebody had, an idea in a very homogenous team of people. {Participant 33, researcher}

Advancing Interdisciplinary Knowledge and Ethical Alignment

Both developers and physicians commented on the need for greater collaboration between stakeholders in AI and ML and medicine, emphasizing that “there’s a big gap in those two communities in terms of the problems that one wants solved, the problems that are solved, how the communication happens, and how that’s all addressed” (participant 12, physicians). This physician noted that “deep collaborations are fairly rare...It’s not easy to find them” [12], in agreement with several researchers who described the current state of research as interdisciplinary on the surface, but still highly localized. Many interviewees perceived a need for greater interdisciplinary knowledge between the 2 fields, emphasizing the desire for “more people who are dual trained: who really deeply understand subject matter and who deeply understand algorithms” (participant 33, researcher), and who can “actually own a scientific question and try to answer it end-to-end” (participant 26, researcher). They felt that a greater commitment to interdisciplinary training and collaboration would lead to the development of more tools with clinical relevance and utility.

Interviewees highlighted how different approaches to ethical and professional standards in AI and ML research and medicine may be sources of conflict when applying AI and ML in medicine. They tended to agree that although research and patient care are united by institutionalized ethical commitments (eg, the Hippocratic Oath, the Belmont Report, and the Common Rule) and organizational safeguards (eg, institutional review boards), ethics in the emerging fields of AI and ML lack institutionalized guidance and typically consist of individual researchers voluntarily following informal guidelines and

recommendations. Several researchers expressed the feeling that “Right now the system relies on people like me doing the right thing.” (participant 10, researcher) and felt that ethics are not adequately prioritized in computer science training and curricula:

We are taught to think, “Here is a thing that your program should do, and then if it does it then you’re good.” But you’re not really taught to think about, what are all the other things that it could be doing as well on the side? We’re just biased towards getting the one positive result out of our program without thinking about all the negative consequences that could happen. {Participant 06, researcher}

Interviewees identified more training in ethics as necessary to support the translation of AI and ML research into robust clinical tools. Several researchers related a desire for enhanced ethical training among developers; one asserted the importance of “incorporating ethical thinking into every single class that computer scientists take, so that it is not just the one throwaway class you have to sit through, but it is like every time you do something, just think about [ethics] as well. Because the whole point is you should...think about the ethics and think about the potential backfiring while you’re designing the technology—not after you’ve designed it” (participant 06, researcher).

Interviewees also perceived a need for increased computer science education in medical training, noting that “there has to be the kind of literacy about computer science that is not currently required in the medical curriculum” (participant 10, researcher). Several physicians expressed similar desires, with one asserting, “We’re going to have to learn something about how these algorithms work...We’re going to own AI just as we’ve kind of owned other kinds of new technologies that have been incorporated into our practice” (participant 14, physician).

Discussion

Background

In this report, we sought the perspectives of researchers and physicians regarding ethical considerations in the translation of ML technologies into medicine. Existing qualitative literature pertaining to medical AI and ML has primarily focused on clinicians’ views on specific uses or implementations of AI and ML in medicine [12,16,22-24,26]. Our study is unique in that its use of open-ended questions allowed interviewees to explore their sentiments and perspectives without overreliance on implicit assumptions about what AI and ML currently are *like*. Because of the open-endedness of the questions, participants articulated the issues that they resonated with most strongly, as opposed to responding to prescriptive questions about issues defined a priori by our research team. To the best of our knowledge, this study is among the first to describe such perspectives.

Our findings revealed a range of ethical concerns shared by both researchers and physicians regarding the initial phase of research, which we have referred to as the “problem formulation” phase or “phase 1” (Figure 1). Although our interview questions did not specifically probe these early issues,

most interviewees discussed them in great detail. Their concerns revolved around several broad themes (ie, influences on research directions, clinical needs and utility, stakeholder involvement, and interdisciplinary knowledge); interviewees viewed themes as interlinked and deserving of critical consideration before the beginning of algorithm development.

Establishing Clinical Need and Value: The Significance of Phase-1 Decision-Making

To date, a small percentage of AI and ML tools developed for use in medicine have been successfully implemented in clinical practice, and for those tools that have been implemented, their acceptability has sometimes been disputed by health care practitioners and administrators [6,7]. For example, clinicians have raised concerns regarding the risks of cognitive burden, overreliance on algorithms, degradation of human clinical abilities, and patient overtreatment in response to several early sepsis detection tools [14-16,42]. Interviewees in our study were aware of this “AI chasm,” and identified processes that take place during phase 1, such as selecting a research question and building a research team, that they felt contribute to these persistent implementation challenges.

Notably, interviewees linked this pattern to a lack of an early and well-defined clinical need, often because of AI and ML development occurring without sufficient input seeking from clinicians, patients, community members, and others. The lack of appropriate stakeholder involvement or the misalignment of values between stakeholders were identified as phase-1 failures that directly contribute to issues in the development and clinical implementation of AI and ML tools, including reduced clinical utility and acceptability. Interviewees agreed that research questions must be sensitive to real-world needs and contextual factors, such as the clinical environments in which health care providers and teams work, and emphasized that these considerations should remain central throughout the full course of development and implementation. These findings align with a qualitative study by Watson et al [30], in which leaders of academic medical centers described identifying a research question as an essential task that must take place before model development begins and suggested that consultation with clinicians and other stakeholders helps greatly in formulating the question [30].

Aligning Values and Motivations: The Tension Between Innovation and Patient Care

Although modern medicine is an established field that prioritizes ethically robust advancement, AI and ML (in their current state) were described as rapidly evolving, technology-centric fields that prioritize innovation. Echoing concerns previously raised in the literature, interviewees described how these divergent priorities may lead to ethical tensions between the individuals and institutions that develop these technologies and the clinicians and patients who ultimately use them [6,21]. Interviewees perceived physicians’ motivations for using medical AI and ML as related to improving patient outcomes and lessening clinical burden, whereas the motivations among developers of medical AI and ML were viewed as more varied and not necessarily aligned with those of the end users.

Notably, a number of researchers agreed that basic, stepwise, or “boring” AI and ML research has benefits that may be undervalued in today’s research culture, in recognition of the understanding that innovation for its own sake is likely not inherently beneficial for the advancement of AI and ML in medicine. These findings reassuringly suggest that the physicians and researchers we interviewed distinguish similarly between the intellectual and moral dimensions of AI and ML research in the health care context, value cautious and measured innovation, and are generally aligned in their understanding that the chief aim of biomedical innovation is to reduce the burden of health disorders and conditions.

Advancing Interdisciplinary Engagement: Recommendations for Strengthening Ethical Innovation in Medicine

Interviewees agreed that medical AI and ML’s success in both the short and long terms will require sustained efforts to engage a broad stakeholder base before development efforts and reimagine interdisciplinary education and training for both developers and clinicians. The value of increased and earlier stakeholder involvement has been previously identified [22,29] and was raised by many interviewees as an actionable strategy for anticipating and meeting current challenges related to problem formulation. Although AI and ML developers possess the technical expertise needed to create algorithms, clinicians possess the insight and professional experience needed to determine how best to integrate a potential tool into an existing clinical space.

As the field of medical AI and ML innovation continues to expand, participants emphasized that it should increasingly involve dual-trained individuals with expertise in both AI or ML and medicine. Expanded opportunities for the dual training of new clinician researchers are greatly needed, in addition to more interdisciplinary training for individuals whose expertise

resides in a single field. This is especially relevant in light of the FDA’s 2022 guidance regarding the 21st Century Cures Act, where it was indicated that clinical decision support software is not classified as a medical device when the health care provider “can independently review the basis for [the] recommendations” [43]. Consequently, AI and ML tools that have logic and inputs that can be reviewed will likely not require the same FDA oversight as other medical devices, shifting the onus of interpreting and verifying the outputs of these tools to clinicians who may have varying levels of understanding of AI and ML technologies. Although there are still many unanswered questions regarding how the FDA’s guidance will affect hospital systems and health care providers as the availability of AI- or ML-enabled clinical tools systems increases, those who have relevant training in AI and ML will be better prepared to understand the functionality of these tools and make confident clinical decisions based on their output [44,45].

Beyond increased technical education, interviewees in this project specifically underscored the need for systematic ethics training and resources for tool developers, with both groups expressing concern regarding the lack of institutionalized ethical standards in the field of AI and ML. They suggest that the lack of ethical consensus within AI and ML may represent a limiting factor for innovation in medicine. This finding indicates that more empirical work is needed to develop a coherent and coordinated framework for reasoning through ethical problems in medical AI and ML, and to develop adequate guidelines, regulations, and safeguards that ensure medical AI and ML’s acceptability to care teams and fulfillment of public trust responsibilities. In working toward greater ethical alignment, interviewees described a myriad of questions related to phase 1 that they felt were important for medical AI and ML teams to consider before the start of algorithm design and development (Table 2).

Table 2. Questions to consider at the start of medical artificial intelligence (AI) or machine learning (ML) projects, as recommended by interviewees.

Need	Questions to consider	Relevant quotes
Assess motivations, priorities, and incentives	What are the motivations for the creation of this tool? Are any of these in conflict with the goals and ideals of the field of medicine? What economic or social incentives may be influencing the motivations?	<ul style="list-style-type: none"> “Am I doing this because I want to innovate or am I doing this because I really want to prove this one process or take care of the patient?” (participant 15, physician)
Involve stakeholders	Have the perspectives of stakeholders who may be affected by the development and use of this tool (patients, family members, clinicians, and hospital staff) been solicited and considered? Have the perspectives of diverse stakeholders been solicited and considered (individuals of different races, ethnicities, genders, sexualities, ages, SES ^a)? Have stakeholders' concerns been addressed and their input incorporated?	<ul style="list-style-type: none"> “Who's asking the questions?” (participant 14, researcher) “Who [hasn't] had the privilege to ask the questions? Who hasn't been empowered to be able to ask the questions?” (participant 33, ML researcher)
Identify problem space	What <i>specific</i> role will this technology fill in medicine? What is the <i>specific</i> problem in medicine that the tool will address? How is this problem currently being addressed? How may it benefit from the use of AI or ML? Has the input of stakeholders been incorporated when identifying the problem space?	<ul style="list-style-type: none"> “What is the problem we're trying to solve? Think of AI as just an enabling technology like anything else...You should never do something for the sake of using AI.” (participant 15, physician)
Evaluate need	Can this problem be solved without AI or ML? Is AI or ML the best tool currently available to address this problem? What are some possible non-AI or non-ML solutions for this problem? Are these more practical, feasible, affordable, accessible? Has the input of stakeholders been incorporated when evaluating the necessity of the AI or ML solution?	<ul style="list-style-type: none"> “Does machine learning actually make the application or the intervention more effective? Do we need to use machine learning?...When does machine learning actually improve things, and when should you maybe not use machine learning or refuse the use of machine learning if it can actually do more harm than good?” (participant 22, physician).
Assess collaborations	Has an interdisciplinary team of collaborators been established? Does the team have the expertise in medicine needed to be able to thoughtfully develop this tool? Will these collaborations be able to continue as the project progresses? Do the collaborators include different types of stakeholders?	<ul style="list-style-type: none"> “[Talk] to different stakeholders to see what things they find as issues, either in the workplace or with their profession, that AI could really assist with. That collaboration...[ensures] that it is something that is actually useful in the medical and healthcare field.” (participant 27, physician)
Push boundaries on interdisciplinary knowledge	What should interdisciplinary knowledge look like? What assumptions about interdisciplinary knowledge and collaboration should be reexamined or challenged in this emerging context?	<ul style="list-style-type: none"> “I cannot stress enough how important it is to have more people who are dual trained: who deeply understand the subject matter and who deeply understand algorithms...Team science is great but in order to do really transformational work, you need some of both on some level.” (participant 33, ML researcher)

^aSES: socioeconomic status.

Limitations

This study had several limitations. Owing to their qualitative nature, the results are not representative; however, the ethical issues raised could be transferable to other similar areas of study in medicine. In addition, because the semistructured design of the interviews emphasized open-ended questions, the ability to compare responses among and between the participant groups was limited. Furthermore, this analysis did not include the perspectives of other potential stakeholder groups, such as patients, ethicists, industry researchers, representatives, or other health care professionals beyond physicians. Additional qualitative and quantitative research is required to confirm these findings. Research involving complementary quantitative approaches could be useful once ethical concerns are articulated, refined, and prioritized. Vignette studies such as surveys that present hypothetical scenarios offer a promising approach to support reproducibility. Future stakeholder studies may benefit

from focusing on the “problem formulation” phase of research, as it presents an early opportunity to avoid costly failures during development and implementation.

Conclusions

In conclusion, this study provides a description of the nuanced views of researchers and physicians regarding ethical considerations in the use of AI and ML technologies in medicine. Considerations related to the earliest processes in a medical AI or ML project, such as selecting a research question and forming a research team, were highlighted by interviewees for their potential to have an outsized impact on the following phases of development and implementation. The phase-1 ethical issues identified by interviewees were notably interdisciplinary in nature and invited questions regarding how to align priorities and values across disciplines and ensure clinical value throughout the development and implementation of medical AI and ML. Relatedly, interviewees suggested interdisciplinary

solutions to these issues, for example, more resources to support knowledge generation and collaboration between developers and physicians, engagement with a broader range of stakeholders, and efforts to increase diversity in research both broadly and within individual teams. Although some of the issues addressed in this paper may be outside the control of any

individual researcher or team, thorough individual- or team-level assessment of these considerations before the development phase could aid in maximizing the benefit of new tools for patients and care teams and ultimately increase the successful uptake of AI and ML innovations.

Acknowledgments

This study was supported by the National Center for Advancing Translational Sciences (R01-TR-003505). The authors acknowledge Kyle McKinley and Jodi Paik for their contributions to this research.

Data Availability

The full data set is not available to protect the identities of the participants. A limited data set is available upon request.

Authors' Contributions

JPK conceptualized the project, led the development of the interview instrument, performed interviews, oversaw qualitative coding and analysis, and contributed to the writing and editing of the manuscript. KR assisted in the development of the interview instrument, led the qualitative coding and analysis, and contributed to the writing and editing of the manuscript. MK performed qualitative coding and analysis and contributed to the writing and editing of the manuscript. JH performed the transcription, participated in qualitative coding and analysis, and contributed to the writing and editing of the manuscript. LWR contributed to the writing and editing of the manuscript. LBD conducted the interviews and contributed to the writing and editing of the manuscript.

Conflicts of Interest

LWR is Editor-in-Chief of Academic Medicine. The other authors declare no competing interests.

References

1. Artificial intelligence and machine learning (AI/ML)-enabled medical devices. U.S. Food and Drug Administration. 2022. URL: <https://www.fda.gov/medical-devices/software-medical-device-samd/artificial-intelligence-and-machine-learning-aiml-enabled-medical-devices> [accessed 2023-03-20]
2. Tang X, Li X, Ding Y, Song M, Bu Y. The pace of artificial intelligence innovations: speed, talent, and trial-and-error. *J Informetr* 2020 Nov;14(4):101094 [FREE Full text] [doi: [10.1016/j.joi.2020.101094](https://doi.org/10.1016/j.joi.2020.101094)]
3. Caballé-Cervigón N, Castillo-Sequera JL, Gómez-Pulido JA, Gómez-Pulido JM, Polo-Luque ML. Machine learning applied to diagnosis of human diseases: a systematic review. *Appl Sci* 2020;10(15):5135 [FREE Full text] [doi: [10.3390/app10155135](https://doi.org/10.3390/app10155135)]
4. Stafford IS, Kellermann M, Mossotto E, Beattie RM, MacArthur BD, Ennis S. A systematic review of the applications of artificial intelligence and machine learning in autoimmune diseases. *NPJ Digit Med* 2020 Mar 9;3(1):30 [FREE Full text] [doi: [10.1038/s41746-020-0229-3](https://doi.org/10.1038/s41746-020-0229-3)] [Medline: [32195365](https://pubmed.ncbi.nlm.nih.gov/32195365/)]
5. Solomonides AE, Koski E, Atabaki SM, Weinberg S, McGreevey JD, Kannry JL, et al. Defining AMIA's artificial intelligence principles. *J Am Med Inform Assoc* 2022 Mar 15;29(4):585-591 [FREE Full text] [doi: [10.1093/jamia/ocac006](https://doi.org/10.1093/jamia/ocac006)] [Medline: [35190824](https://pubmed.ncbi.nlm.nih.gov/35190824/)]
6. McCradden MD, Anderson JA, A Stephenson E, Drysdale E, Erdman L, Goldenberg A, et al. A research ethics framework for the clinical translation of healthcare machine learning. *Am J Bioeth* 2022 May 20;22(5):8-22. [doi: [10.1080/15265161.2021.2013977](https://doi.org/10.1080/15265161.2021.2013977)] [Medline: [35048782](https://pubmed.ncbi.nlm.nih.gov/35048782/)]
7. Peterson ED. Machine learning, predictive analytics, and clinical practice: can the past inform the present? *JAMA* 2019 Dec 17;322(23):2283-2284. [doi: [10.1001/jama.2019.17831](https://doi.org/10.1001/jama.2019.17831)] [Medline: [31755902](https://pubmed.ncbi.nlm.nih.gov/31755902/)]
8. Earl J. Innovative practice, clinical research, and the ethical advancement of medicine. *Am J Bioeth* 2019 Jun 28;19(6):7-18 [FREE Full text] [doi: [10.1080/15265161.2019.1602175](https://doi.org/10.1080/15265161.2019.1602175)] [Medline: [31135322](https://pubmed.ncbi.nlm.nih.gov/31135322/)]
9. Quinn T, Senadeera M, Jacobs S, Coghlan S, Le V. Trust and medical AI: the challenges we face and the expertise needed to overcome them. *J Am Med Inform Assoc* 2021 Mar 18;28(4):890-894 [FREE Full text] [doi: [10.1093/jamia/ocaa268](https://doi.org/10.1093/jamia/ocaa268)] [Medline: [33340404](https://pubmed.ncbi.nlm.nih.gov/33340404/)]
10. Barak-Corren Y, Chaudhari P, Perniciaro J, Waltzman M, Fine AM, Reis BY. Prediction across healthcare settings: a case study in predicting emergency department disposition. *NPJ Digit Med* 2021 Dec 15;4(1):169 [FREE Full text] [doi: [10.1038/s41746-021-00537-x](https://doi.org/10.1038/s41746-021-00537-x)] [Medline: [34912043](https://pubmed.ncbi.nlm.nih.gov/34912043/)]
11. McDermott MB, Wang S, Marinsek N, Ranganath R, Foschini L, Ghassemi M. Reproducibility in machine learning for health research: still a ways to go. *Sci Transl Med* 2021 Mar 24;13(586):eabb1655. [doi: [10.1126/scitranslmed.abb1655](https://doi.org/10.1126/scitranslmed.abb1655)] [Medline: [33762434](https://pubmed.ncbi.nlm.nih.gov/33762434/)]

12. Henry KE, Kornfield R, Sridharan A, Linton RC, Groh C, Wang T, et al. Human-machine teaming is key to AI adoption: clinicians' experiences with a deployed machine learning system. *NPJ Digit Med* 2022 Jul 21;5(1):97 [FREE Full text] [doi: [10.1038/s41746-022-00597-7](https://doi.org/10.1038/s41746-022-00597-7)] [Medline: [35864312](https://pubmed.ncbi.nlm.nih.gov/35864312/)]
13. Obermeyer Z, Powers B, Vogeli C, Mullainathan S. Dissecting racial bias in an algorithm used to manage the health of populations. *Science* 2019 Oct 25;366(6464):447-453. [doi: [10.1126/science.aax2342](https://doi.org/10.1126/science.aax2342)] [Medline: [31649194](https://pubmed.ncbi.nlm.nih.gov/31649194/)]
14. Elish MC. The stakes of uncertainty: developing and integrating machine learning in clinical care. *Ethnograph Praxis Indust Conf Proc* 2019 Jan 24;2018(1):364-380 [FREE Full text] [doi: [10.1111/1559-8918.2018.01213](https://doi.org/10.1111/1559-8918.2018.01213)]
15. Joshi M, Mecklai K, Rozenblum R, Samal L. Implementation approaches and barriers for rule-based and machine learning-based sepsis risk prediction tools: a qualitative study. *JAMIA Open* 2022 Jul;5(2):ooac022 [FREE Full text] [doi: [10.1093/jamiaopen/ooac022](https://doi.org/10.1093/jamiaopen/ooac022)] [Medline: [35474719](https://pubmed.ncbi.nlm.nih.gov/35474719/)]
16. Sandhu S, Lin AL, Brajer N, Sperling J, Ratliff W, Bedoya AD, et al. Integrating a machine learning system into clinical workflows: qualitative study. *J Med Internet Res* 2020 Nov 19;22(11):e22421 [FREE Full text] [doi: [10.2196/22421](https://doi.org/10.2196/22421)] [Medline: [33211015](https://pubmed.ncbi.nlm.nih.gov/33211015/)]
17. Char D, Abramoff M, Feudtner C. A framework to evaluate ethical considerations with ML-HCA applications-valuable, even necessary, but never comprehensive. *Am J Bioeth* 2020 Nov;20(11):W6-10 [FREE Full text] [doi: [10.1080/15265161.2020.1827695](https://doi.org/10.1080/15265161.2020.1827695)] [Medline: [33103985](https://pubmed.ncbi.nlm.nih.gov/33103985/)]
18. Andersen TO, Nunes F, Wilcox L, Kazianus E, Matthiesen S, Magrabi F. Realizing AI in healthcare: challenges appearing in the wild. In: proceedings of the Extended Abstracts of the 2021 CHI Conference on Human Factors in Computing Systems. 2021 May Presented at: CHI EA '21; May 8-13, 2021; Yokohama, Japan p. 1-5 URL: <https://dl.acm.org/doi/10.1145/3411763.3441347> [doi: [10.1145/3411763.3441347](https://doi.org/10.1145/3411763.3441347)]
19. Bridge to artificial intelligence (Bridge2AI). National Institutes of Health. URL: <https://commonfund.nih.gov/bridge2ai> [accessed 2023-03-20]
20. Green J, Britten N. Qualitative research and evidence based medicine. *BMJ* 1998 Apr 18;316(7139):1230-1232 [FREE Full text] [doi: [10.1136/bmj.316.7139.1230](https://doi.org/10.1136/bmj.316.7139.1230)] [Medline: [9583929](https://pubmed.ncbi.nlm.nih.gov/9583929/)]
21. Char DS, Shah NH, Magnus D. Implementing machine learning in health care - addressing ethical challenges. *N Engl J Med* 2018 Mar 15;378(11):981-983 [FREE Full text] [doi: [10.1056/NEJMp1714229](https://doi.org/10.1056/NEJMp1714229)] [Medline: [29539284](https://pubmed.ncbi.nlm.nih.gov/29539284/)]
22. McCradden MD, Baba A, Saha A, Ahmad S, Boparai K, Fadaiefard P, et al. Ethical concerns around use of artificial intelligence in health care research from the perspective of patients with meningioma, caregivers and health care providers: a qualitative study. *CMAJ Open* 2020 Feb 18;8(1):E90-E95 [FREE Full text] [doi: [10.9778/cmajo.20190151](https://doi.org/10.9778/cmajo.20190151)] [Medline: [32071143](https://pubmed.ncbi.nlm.nih.gov/32071143/)]
23. Parikh RB, Manz CR, Nelson MN, Evans CN, Regli SH, O'Connor N, et al. Clinician perspectives on machine learning prognostic algorithms in the routine care of patients with cancer: a qualitative study. *Support Care Cancer* 2022 May;30(5):4363-4372 [FREE Full text] [doi: [10.1007/s00520-021-06774-w](https://doi.org/10.1007/s00520-021-06774-w)] [Medline: [35094138](https://pubmed.ncbi.nlm.nih.gov/35094138/)]
24. Benda NC, Das LT, Abramson EL, Blackburn K, Thoman A, Kaushal R, et al. "How did you get to this number?" Stakeholder needs for implementing predictive analytics: a pre-implementation qualitative study. *J Am Med Inform Assoc* 2020 May 01;27(5):709-716 [FREE Full text] [doi: [10.1093/jamia/ocaa021](https://doi.org/10.1093/jamia/ocaa021)] [Medline: [32159774](https://pubmed.ncbi.nlm.nih.gov/32159774/)]
25. Jacobs M, He J, Pradier MF, Lam B, Ahn AC, McCoy TH, et al. Designing AI for trust and collaboration in time-constrained medical decisions: a sociotechnical lens. In: Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems. 2021 Presented at: CHI '21; May 8-13, 2021; Yokohama, Japan p. 1-14 URL: <https://dl.acm.org/doi/10.1145/3411764.3445385> [doi: [10.1145/3411764.3445385](https://doi.org/10.1145/3411764.3445385)]
26. Poncette AS, Spies C, Mosch L, Schieler M, Weber-Carstens S, Krampe H, et al. Clinical requirements of future patient monitoring in the intensive care unit: qualitative study. *JMIR Med Inform* 2019 Apr 30;7(2):e13064 [FREE Full text] [doi: [10.2196/13064](https://doi.org/10.2196/13064)] [Medline: [31038467](https://pubmed.ncbi.nlm.nih.gov/31038467/)]
27. Tang L, Li J, Fantus S. Medical artificial intelligence ethics: a systematic review of empirical studies. *Digit Health* 2023 Jul 06;9:20552076231186064 [FREE Full text] [doi: [10.1177/20552076231186064](https://doi.org/10.1177/20552076231186064)] [Medline: [37434728](https://pubmed.ncbi.nlm.nih.gov/37434728/)]
28. Park Y, Jackson GP, Foreman MA, Gruen D, Hu J, Das AK. Evaluating artificial intelligence in medicine: phases of clinical research. *JAMIA Open* 2020 Sep 08;3(3):326-331 [FREE Full text] [doi: [10.1093/jamiaopen/ooaa033](https://doi.org/10.1093/jamiaopen/ooaa033)] [Medline: [33215066](https://pubmed.ncbi.nlm.nih.gov/33215066/)]
29. Elish MC, Watkins EA. Repairing innovation: a study of integrating AI in clinical care. *Data & Society*. 2020. URL: <https://datasociety.net/wp-content/uploads/2020/09/Repairing-Innovation-DataSociety-20200930-1.pdf> [accessed 2023-10-06]
30. Watson J, Hutryra CA, Clancy SM, Chandiramani A, Bedoya A, Ilangovan K, et al. Overcoming barriers to the adoption and implementation of predictive modeling and machine learning in clinical care: what can we learn from US academic medical centers? *JAMIA Open* 2020 Jul;3(2):167-172 [FREE Full text] [doi: [10.1093/jamiaopen/ooz046](https://doi.org/10.1093/jamiaopen/ooz046)] [Medline: [32734155](https://pubmed.ncbi.nlm.nih.gov/32734155/)]
31. Seneviratne MG, Shah NH, Chu L. Bridging the implementation gap of machine learning in healthcare. *BMJ Innov* 2019 Dec 20;6(2):45-47 [FREE Full text] [doi: [10.1136/bmjinnov-2019-000359](https://doi.org/10.1136/bmjinnov-2019-000359)]
32. Kim JP. Letter to the editor: machine learning and artificial intelligence in psychiatry: balancing promise and reality. *J Psychiatr Res* 2021 Apr;136:244-245. [doi: [10.1016/j.jpsychires.2021.02.021](https://doi.org/10.1016/j.jpsychires.2021.02.021)] [Medline: [33621909](https://pubmed.ncbi.nlm.nih.gov/33621909/)]
33. Sandelowski M. Whatever happened to qualitative description? *Res Nurs Health* 2000 Aug;23(4):334-340. [doi: [10.1002/1098-240x\(200008\)23:4<334::aid-nur9>3.0.co;2-g](https://doi.org/10.1002/1098-240x(200008)23:4<334::aid-nur9>3.0.co;2-g)] [Medline: [10940958](https://pubmed.ncbi.nlm.nih.gov/10940958/)]

34. Neergaard MA, Olesen F, Andersen RS, Sondergaard J. Qualitative description - the poor cousin of health research? *BMC Med Res Methodol* 2009 Jul 16;9:52 [FREE Full text] [doi: [10.1186/1471-2288-9-52](https://doi.org/10.1186/1471-2288-9-52)] [Medline: [19607668](https://pubmed.ncbi.nlm.nih.gov/19607668/)]
35. Kim H, Sefcik JS, Bradway C. Characteristics of qualitative descriptive studies: a systematic review. *Res Nurs Health* 2017 Feb;40(1):23-42 [FREE Full text] [doi: [10.1002/nur.21768](https://doi.org/10.1002/nur.21768)] [Medline: [27686751](https://pubmed.ncbi.nlm.nih.gov/27686751/)]
36. Palinkas LA, Horwitz SM, Green CA, Wisdom JP, Duan N, Hoagwood K. Purposeful sampling for qualitative data collection and analysis in mixed method implementation research. *Adm Policy Ment Health* 2015 Sep;42(5):533-544 [FREE Full text] [doi: [10.1007/s10488-013-0528-y](https://doi.org/10.1007/s10488-013-0528-y)] [Medline: [24193818](https://pubmed.ncbi.nlm.nih.gov/24193818/)]
37. Saunders B, Sim J, Kingstone T, Baker S, Waterfield J, Bartlam B, et al. Saturation in qualitative research: exploring its conceptualization and operationalization. *Qual Quant* 2018 Mar;52(4):1893-1907 [FREE Full text] [doi: [10.1007/s11135-017-0574-8](https://doi.org/10.1007/s11135-017-0574-8)] [Medline: [29937585](https://pubmed.ncbi.nlm.nih.gov/29937585/)]
38. Morse JM. Designing funded qualitative research. In: Denzin NK, Lincoln YS, editors. *Handbook of Qualitative Research*. Thousand Oaks, CA: Sage Publications; 1994:220-235.
39. Miles M, Huberman A. *Qualitative Data Analysis: An Expanded Sourcebook*. 2nd edition. Thousand Oaks, CA: Sage Publications; 1994.
40. Hsieh H, Shannon SE. Three approaches to qualitative content analysis. *Qual Health Res* 2005 Nov;15(9):1277-1288. [doi: [10.1177/1049732305276687](https://doi.org/10.1177/1049732305276687)] [Medline: [16204405](https://pubmed.ncbi.nlm.nih.gov/16204405/)]
41. Downe-Wamboldt B. Content analysis: method, applications, and issues. *Health Care Women Int* 1992 Jul;13(3):313-321. [doi: [10.1080/07399339209516006](https://doi.org/10.1080/07399339209516006)] [Medline: [1399871](https://pubmed.ncbi.nlm.nih.gov/1399871/)]
42. Adams R, Henry KE, Sridharan A, Soleimani H, Zhan A, Rawat N, et al. Prospective, multi-site study of patient outcomes after implementation of the TREWS machine learning-based early warning system for sepsis. *Nat Med* 2022 Jul;28(7):1455-1460. [doi: [10.1038/s41591-022-01894-0](https://doi.org/10.1038/s41591-022-01894-0)] [Medline: [35864252](https://pubmed.ncbi.nlm.nih.gov/35864252/)]
43. Clinical decision support: guidance for industry and food and drug administration staff. U.S. Food and Drug Administration. URL: <https://www.fda.gov/media/109618/download> [accessed 2023-07-19]
44. Weissman GE. FDA regulation of predictive clinical decision-support tools: what does it mean for hospitals? *J Hosp Med* 2021 Apr 19;16(4):244-246 [FREE Full text] [doi: [10.12788/jhm.3450](https://doi.org/10.12788/jhm.3450)] [Medline: [32853146](https://pubmed.ncbi.nlm.nih.gov/32853146/)]
45. Jackups JR. FDA regulation of laboratory clinical decision support software: is it a medical device? *Clin Chem* 2023 Apr 03;69(4):327-329. [doi: [10.1093/clinchem/hvad011](https://doi.org/10.1093/clinchem/hvad011)] [Medline: [36806588](https://pubmed.ncbi.nlm.nih.gov/36806588/)]

Abbreviations

AI: artificial intelligence

FDA: Food and Drug Administration

ML: machine learning

Edited by K El Emam, B Malin; submitted 21.03.23; peer-reviewed by R Hendricks-Sturup, C Lai; comments to author 25.06.23; revised version received 20.08.23; accepted 16.09.23; published 30.10.23.

Please cite as:

Kim JP, Ryan K, Kasun M, Hogg J, Dunn LB, Roberts LW

Physicians' and Machine Learning Researchers' Perspectives on Ethical Issues in the Early Development of Clinical Machine Learning Tools: Qualitative Interview Study

JMIR AI 2023;2:e47449

URL: <https://ai.jmir.org/2023/1/e47449>

doi: [10.2196/47449](https://doi.org/10.2196/47449)

PMID: [38875536](https://pubmed.ncbi.nlm.nih.gov/38875536/)

©Jane Paik Kim, Katie Ryan, Max Kasun, Justin Hogg, Laura B Dunn, Laura Weiss Roberts. Originally published in *JMIR AI* (<https://ai.jmir.org>), 30.10.2023. This is an open-access article distributed under the terms of the Creative Commons Attribution License (<https://creativecommons.org/licenses/by/4.0/>), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in *JMIR AI*, is properly cited. The complete bibliographic information, a link to the original publication on <https://www.ai.jmir.org/>, as well as this copyright and license information must be included.

Original Paper

Real-Time Classification of Causes of Death Using AI: Sensitivity Analysis

Patrícia Pita Ferreira^{1,2,3}, MD; Diogo Godinho Simões^{1,4}, MD; Constança Pinto de Carvalho^{1,5}, MD; Francisco Duarte⁶, MSc; Eugénia Fernandes¹, PhD; Pedro Casaca Carvalho¹, MD; José Francisco Loff⁷, MSc; Ana Paula Soares¹, MPH; Maria João Albuquerque¹, BSc; Pedro Pinto-Leite¹, MD; André Peralta-Santos^{1,8,9}, MPH, MD, PhD

¹Direção de Serviços de Informação e Análise, Direção-Geral da Saúde, Lisbon, Portugal

²Unidade de Saúde Pública Zé Povinho, Agrupamento de Centros de Saúde do Oeste Norte, Administração Regional de Saúde de Lisboa e Vale do Tejo, Caldas da Rainha, Portugal

³NOVA National School of Public Health, Universidade NOVA de Lisboa, Lisbon, Portugal

⁴Unidade de Saúde Pública Higeia, Agrupamento de Centros de Saúde de Almada-Seixal, Administração Regional de Saúde de Lisboa e Vale do Tejo, Almada, Portugal

⁵Unidade de Saúde Pública do Litoral Alentejano, Unidade Local de Saúde do Litoral Alentejano, Administração Regional de Saúde do Alentejo, Santiago do Cacém, Portugal

⁶Lisbon, Portugal

⁷phiStat – Statistical Consulting, Lisbon, Portugal

⁸NOVA National School of Public Health, Public Health Research Centre, Universidade NOVA de Lisboa, Lisbon, Portugal

⁹Comprehensive Health Research Centre, Universidade NOVA de Lisboa, Lisbon, Portugal

Corresponding Author:

Patrícia Pita Ferreira, MD

Direção de Serviços de Informação e Análise

Direção-Geral da Saúde

Alameda D. Afonso Henriques, 45

Lisbon, 1049-005

Portugal

Phone: 351 218430500

Email: ppita.ferreira@gmail.com

Abstract

Background: In 2021, the European Union reported >270,000 excess deaths, including >16,000 in Portugal. The Portuguese Directorate-General of Health developed a deep neural network, AUTOCOD, which determines the primary causes of death by analyzing the free text of physicians' death certificates (DCs). Although AUTOCOD's performance has been established, it remains unclear whether its performance remains consistent over time, particularly during periods of excess mortality.

Objective: This study aims to assess the sensitivity and other performance metrics of AUTOCOD in classifying underlying causes of death compared with manual coding to identify specific causes of death during periods of excess mortality.

Methods: We included all DCs between 2016 and 2019. AUTOCOD's performance was evaluated by calculating various performance metrics, such as sensitivity, specificity, positive predictive value (PPV), and F_1 -score, using a confusion matrix. This compared *International Statistical Classification of Diseases and Health-Related Problems, 10th Revision* (ICD-10), classifications of DCs by AUTOCOD with those by human coders at the Directorate-General of Health (gold standard). Subsequently, we compared periods without excess mortality with periods of excess, severe, and extreme excess mortality. We defined excess mortality as 2 consecutive days with a Z score above the 95% baseline limit, severe excess mortality as 2 consecutive days with a Z score >4 SDs, and extreme excess mortality as 2 consecutive days with a Z score >6 SDs. Finally, we repeated the analyses for the 3 most common ICD-10 chapters focusing on block-level classification.

Results: We analyzed a large data set comprising 330,098 DCs classified by both human coders and AUTOCOD. AUTOCOD demonstrated high sensitivity (≥ 0.75) for 10 ICD-10 chapters examined, with values surpassing 0.90 for the more prevalent chapters (chapter II—"Neoplasms," chapter IX—"Diseases of the circulatory system," and chapter X—"Diseases of the respiratory system"), accounting for 67.69% (223,459/330,098) of all human-coded causes of death. No substantial differences were observed in these high-sensitivity values when comparing periods without excess mortality with periods of excess, severe, and extreme

excess mortality. The same holds for specificity, which exceeded 0.96 for all chapters examined, and for PPV, which surpassed 0.75 in 9 chapters, including the more prevalent ones. When considering block classification within the 3 most common ICD-10 chapters, AUTOCOD maintained a high performance, demonstrating high sensitivity (≥ 0.75) for 13 ICD-10 blocks, high PPV for 9 blocks, and specificity of >0.98 in all blocks, with no significant differences between periods without excess mortality and those with excess mortality.

Conclusions: Our findings indicate that, during periods of excess and extreme excess mortality, AUTOCOD's performance remains unaffected by potential text quality degradation because of pressure on health services. Consequently, AUTOCOD can be dependably used for real-time cause-specific mortality surveillance even in extreme excess mortality situations.

(JMIR AI 2023;2:e40965) doi:[10.2196/40965](https://doi.org/10.2196/40965)

KEYWORDS

artificial intelligence; AI; mortality; deep neural networks; evaluation; machine learning; deep learning; mortality statistics; underlying cause of death

Introduction

Background

In 2021, over 270,000 excess deaths were registered in the European Union, with $>16,000$ attributable to Portugal [1]. Although most of these excess deaths were possibly related to the COVID-19 pandemic, excess deaths are generally attributable to preventable causes, making a case for the importance of real-time cause-specific mortality surveillance and the subsequent timely and appropriate public health response and suitable health policies in periods of excess mortality [2].

The Portuguese Directorate-General of Health (DGS) is responsible for processing data from the Death Certificate Information System (SICO) and ensuring the epidemiological surveillance of mortality [3]. SICO all-cause mortality data are automatically analyzed and can be publicly accessed [4]. However, the analysis of death certificates (DCs) requires manual coding of the primary causes of death according to the *International Statistical Classification of Diseases and Health-Related Problems, 10th Revision (ICD-10)* [5]. This manual coding is a resource-intensive task that hinders real-time cause-specific mortality surveillance.

Excess mortality is defined by the World Health Organization as mortality above what would be expected. It allows for assessing the magnitude of a potential public health crisis by checking the additional deaths compared with a reference period and subsequently analyzing their causes in depth [6,7].

Excess mortality can be estimated in several ways. In Portugal, a period of excess mortality is defined as a consecutive period starting with 2 observed numbers of deaths above the baseline's upper 95% confidence limit or with only 1 observed number of deaths above the upper 99% confidence limit of the baseline. The period ends with 2 consecutive values below this limit [8]. This methodology is aligned with the practice of the European mortality monitoring project (EuroMOMO), which allows for the detection and measurement in real time of periods of excess mortality from all causes as a result of threats to public health in Europe [9].

Most excess mortality surveillance systems such as EuroMOMO or national systems are based on all-cause mortality surveillance to ensure real-time surveillance. However, in many countries,

information on cause of death is not readily available as it requires a human step to code the basic cause of death, delaying the surveillance and monitoring of cause-specific mortality. For instance, in Portugal, the manual establishment of the primary causes of death for the previous year is completed by March of the following year [10,11].

To overcome this problem, Portugal developed a deep neural network called AUTOCOD [12,13], which allows for presuggesting primary causes of mortality based on historical data of DCs (except for neonatal and perinatal mortality), achieving accuracies of 89% and 81% for ICD-10 chapters and blocks, respectively. AUTOCOD can also analyze data from autopsy reports and clinical bulletins (deaths occurring in health care facilities). Ultimately, the developed algorithm increased the productivity of coders, sped up the issuance of results and information, and ensured near-real-time mortality surveillance [12,13].

To our knowledge, no widespread dissemination of complex artificial intelligence (AI) algorithms can suggest underlying causes of death through free-text analysis of DCs in the same way as AUTOCOD [14].

Objectives

This study aimed to determine the sensitivity and specificity of AUTOCOD for classifying the underlying cause of death compared with manual coding to ascertain the specific causes of death in periods of excess mortality.

AUTOCOD has already proven to have high sensitivity, specificity, and accuracy in periods without excess mortality. However, it was still being determined whether this performance would be maintained in periods of excess mortality, in which the recording of free text in DCs could change owing to the pressure felt in health services and the need to respond to more requests for DCs. A satisfactory performance by AUTOCOD could pave the way for its implementation as a real-time surveillance tool to monitor cause-specific mortality even during periods in which the national health system experiences severe pressure [14,15].

Methods

Study Population

In this study, we included all DCs registered in Portugal's SICO starting from January 1, 2016, to August 8, 2019. We excluded DCs referring to neonatal, perinatal, and maternal mortality as the AUTOCOD algorithm is not trained for these underlying causes of death [13]. Each DC was manually classified according to the ICD-10 by human coders at the DGS (gold standard) or automatically by AUTOCOD.

Study Design and Data Sets

The methods behind the construction of the AUTOCOD algorithm have been explained in detail in previous publications. The algorithm was initially trained and tested using a data set different from the one chosen for this study [12,13]. The manual codification of causes of death adheres to the World Health Organization Nomenclature Regulations specified in the ICD-10. In addition, it uses the ICD-10 rules for selecting the underlying cause of death as the primary cause of death by international rules [5].

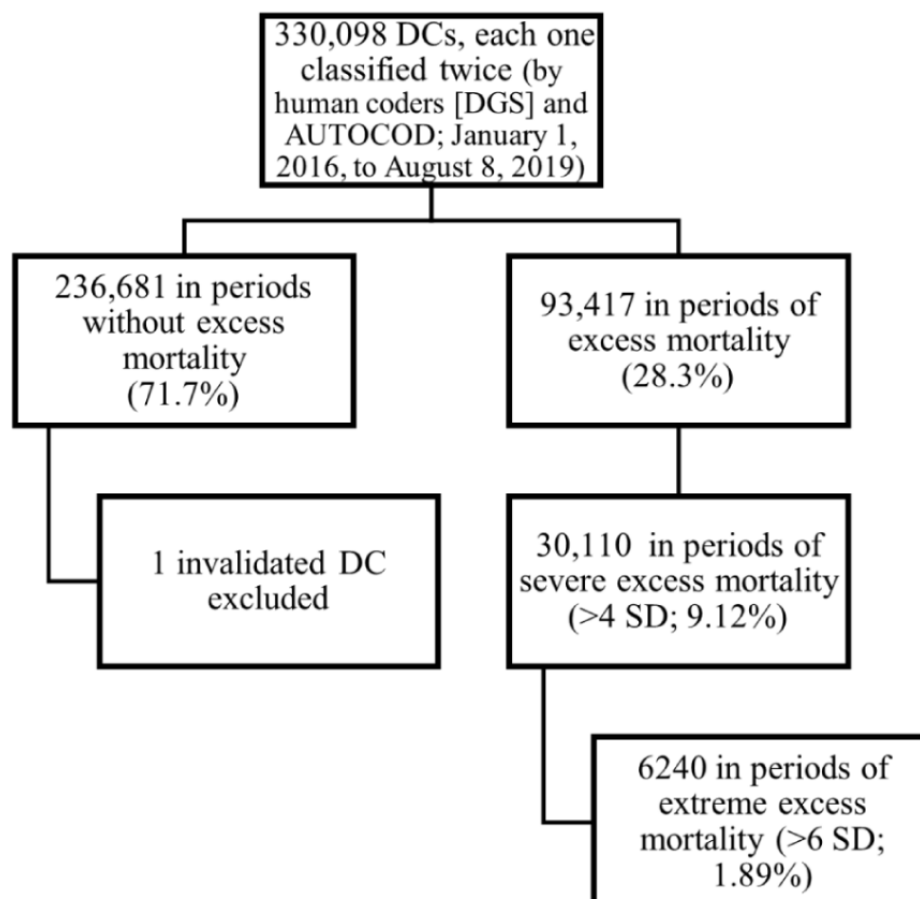
The DC data set was then linked with 2 dictionaries of the ICD-10 to translate block and chapter codes into text descriptions. The DC data set was also linked to the national surveillance all-cause mortality data set [4], which defines the baseline for expected deaths according to the EuroMOMO methodology [16] and the daily count of observed deaths.

Excess Mortality Definition

Using this data set, we defined the periods in which excess mortality was observed according to the EuroMOMO Z score for excess mortality and the rules of Westgard [17] (ie, we considered excess mortality when there were 2 consecutive days with a Z score above the limit at 95% of the baseline or just 1 day at >99%). The period of excess mortality ended with 2 consecutive days below the limit of 95% of the baseline. Flowchart of the study population inclusion criteria can be found in Figure 1.

We also defined 2 metrics for periods of severe and extreme excess mortality. These were 2 consecutive days with a Z score above the limit of 4 SDs and 6 SDs, respectively. The Westgard functions used to classify the different periods can be found in Multimedia Appendix 1 [17-19].

Figure 1. Flowchart of the study population inclusion criteria. DC: death certificate; DGS: Directorate-General of Health.



Statistical Analysis

To obtain the multiclass confusion matrix, we used the “confusionMatrix” function of the *caret* package in RStudio (version 6.0-90; Posit, PBC) [18,19]. In a multiclass problem such as classifying ICD-10 chapters and blocks, the

“confusionMatrix” will show a set of “one-versus-all” results. For example, in a 3-class problem, the sensitivity of the first class is calculated against all the samples in the second and third classes (and so on). The resulting confusion matrix summarizes the prediction results for a classification problem.

The number of correct and incorrect predictions is summarized with count values and broken down by each class. The confusion matrix shows how a classification model such as AUTOCOD is confused when it makes predictions. These numbers are then organized into a table or matrix. Each row of the matrix corresponds to a predicted class (ie, AUTOCOD). Each matrix column corresponds to an actual class (ie, human coders at the DGS).

The numbers of correct and incorrect classifications are then filled into the table. The total number of correct predictions for a class goes into the expected row for that class value and the predicted column for that class value. In the same way, the total number of incorrect predictions for a class goes into the expected row for that class value and the predicted column for that class value.

Finally, we performed a sensitivity analysis (also using the R package *caret*) to compare the classification results obtained using the AUTOCOD algorithm (index test) with the classification made by human coders (gold standard) [20]. This allowed us to obtain the number of true positives and false positives as well as additional metrics such as sensitivity (recall), specificity, accuracy, positive predictive value (PPV), and F_1 -score [13]. This step was performed over time, including a comparison between periods of excess and no excess mortality and between periods of extreme excess mortality and no excess mortality both by chapter and block classification levels of the ICD-10 [13]. We present this comparison as the difference in absolute values and with the Kullback-Leibler divergence (KLD), which measures the distribution of a metric and chapter or block during a specific period of excess or extreme mortality and periods of no excess mortality. In other words, the KLD measures the difference between 2 probability distributions. We used the `kullback_leibler_distanc` function of the R package *philentropy* [21].

The formulas used for all these performance metrics can be found in Table S1 in [Multimedia Appendix 1](#) [17-19].

To assess the quality of AUTOCOD, we opted to present the weighted average of performance metrics such as sensitivity, precision, and F_1 -scores by taking the mean of all class performance metrics while considering each class's number of actual occurrences in the data set. The "weight" refers to the proportion of each class's actual occurrences in the data set relative to the sum of all occurrences. The full formula for this calculation of the weighted average is provided in [Multimedia Appendix 1](#) [17-19]. This choice was made as opposed to presenting the macroaverage of performance metrics (ie, macroaverages assign equal importance to each chapter or block, thus calculating the arithmetic mean of performance metrics) [13] as the latter methodology would artificially increase the importance of the average of the rare or infrequent cause of death chapters and blocks.

In the data set, 1 DC was not adequately codified by AUTOCOD, so the ICD-10 classifications of that DC from both AUTOCOD and the DGS were excluded.

All analyses were performed using R statistical software (version 4.1.2; R Foundation for Statistical Computing) [22-25]. The analyses were checked by 2 researchers.

Ethical Considerations

The DGS is the national entity responsible for data treatment and data protection of the SICO. The data provided were only for the purposes strictly necessary for this study within the competencies of the DGS. Data were previously anonymized. Patient consent was waived as the data were deidentified and processed for reasons of public interest in public health. This research received previous authorization from the DGS following positive advice from its data protection officer. In this way, the research complies with the best practices of the General Data Protection Regulation. This study was exempt from an ethics review board assessment following the self-assessment checklist for ethics of the Ethics Committee of the National School of Public Health [26].

Results

Description of the Data Set

The data set ([Table 1](#)) comprised 330,098 DCs, each classified twice, meaning that we had all DCs classified by human coders and by AUTOCOD. The 3 most common ICD-10 chapters classified by human coders were chapter IX—"Diseases of the circulatory system" (97,420/330,098, 29.51%), chapter II—"Neoplasms" (85,837/330,098, 26%), and chapter X—"Diseases of the respiratory system" (40,202/330,098, 12.18%). A more extensive and detailed descriptive analysis of this data set can be found in [Multimedia Appendix 1](#) [17-19], including the desegregation of DCs by year, ICD-10 chapter or block, and period.

As expected, there were fewer DCs for periods of excess mortality ($n=186,834$; 93,417/330,098, 28.3% of the total DCs from each source) than for periods without excess mortality ($n=473,362$; 236,681/330,098, 71.7% of the total DCs from each source). When considering the periods of severe and extreme excess mortality either for Z scores of >4 SDs ($n=60,220$; 30,110/330,098, 9.12% from each source) or Z scores of >6 SDs ($n=12,480$; 6,240/330,098, 1.89% from each source), the DCs were even fewer.

Considering only the 3 most common chapters of the data set (chapters II, IX, and X), we performed the same analysis for the classification of ICD-10 blocks ([Table 2](#)), which accounted for 67.69% (223,459/330,098) of the total DCs throughout the period. The 5 most common blocks classified in DCs were C00-C97 (malignant neoplasms), I60-I69 (cerebrovascular diseases), I30-I52 (other forms of heart disease), I20-I25 (ischemic heart disease), and J09-J18 (influenza and pneumonia).

Table 1. Description of the study population by excess mortality and type of death certificate coding (N=330,098)^a.

Chapter	No excess mortality, n/N (%)		Excess mortality, n/N (%)		Severe excess mortality (>4 SDs), n/N (%)		Extreme excess mortality (>6 SDs), n/N (%)		Total, n/N (%)	
	Human	AUTOCOD	Human	AU-TOCOD	Human	AU-TOCOD	Human	AU-TOCOD	Human	AUTOCOD
I ^b	4460/ 6154 (72.45)	4863/ 6649 (73.14)	1696/ 6156 (27.55)	1786/6649 (26.86)	546/ 6154 (8.87)	566/ 6649 (8.51)	111/ 6154 (1.8)	117/ 6649 (1.76)	6156/ 330,098 (1.86)	6649/ 330,098 (2.01)
II ^c	64,701/ 85,837 (75.38)	62,895/ 83,462 (75.36)	21,136/ 85,837 (24.62)	20,567/ 83,462 (24.64)	6088/ 85,837 (7.09)	5941/ 83,462 (7.12)	1166/ 85,837 (1.36)	1162/ 83,462 (1.39)	85,837/ 330,098 (26)	83,462/ 330,098 (25.28)
III ^d	916/ 1334 (68.67)	1152/ 1602 (71.91)	418/ 1334 (31.33)	450/ 1602 (28.09)	139/ 1334 (10.42)	150/ 1602 (9.36)	22/ 1334 (1.65)	25/ 1602 (1.56)	1334/ 330,098 (0.4)	1602/ 330,098 (0.49)
IV ^e	11,637/ 16,430 (70.83)	13,727/ 19,382 (70.82)	4793/ 16,430 (29.17)	5655/ 19,382 (29.18)	1594/ 16,430 (9.7)	1880/ 19,382 (9.7)	313/ 16,430 (1.91)	374/ 19,382 (1.93)	16,430/ 330,098 (4.98)	19,382/ 330,098 (5.87)
V ^f	8986/ 12,742 (70.52)	8512/ 12,172 (69.93)	3756/ 12,742 (29.48)	3660/ 12,172 (30.07)	1264/ 12,742 (9.92)	1221/ 12,172 (10.03)	281/ 12,742 (2.21)	261/ 12,172 (2.14)	12,742/ 330,098 (3.86)	12,172/ 330,098 (3.69)
VI ^g	8354/ 11,810 (70.74)	7757/ 10,997 (70.54)	3456/ 11,810 (29.26)	3240/ 10,997 (29.46)	1097/ 11,810 (9.29)	1024/ 10,997 (9.31)	254/ 11,810 (2.15)	228/ 10,997 (2.07)	11,810/ 330,098 (3.58)	10,997/ 330,098 (3.33)
VII ^h	1/2 (50)	— ⁱ	1/2 (50)	—	0/2 (0)	—	0/2 (0)	—	2/330,098 (0)	0/330,098 (0)
VIII ^j	22/30 (73)	6/9 (66.67)	8/30 (26.67)	3/9 (33.33)	5/30 (16.67)	3/9 (33.33)	0/30 (0)	0/9 (0)	30/330,098 (0.01)	9/330,098 (0)
IX ^k	69,021/ 97,420 (70.85)	68,850/ 97,252 (70.8)	28,399/ 97,420 (29.15)	28,402/ 97,252 (29.2)	9287/ 97,420 (9.53)	9296/ 97,252 (9.56)	1918/ 97,420 (1.97)	1937/ 97,252 (1.99)	97,420/ 330,098 (29.51)	97,252/ 330,098 (29.46)
X ^l	26,736/ 40,202 (66.5)	28,913/ 43,057 (67.15)	13,466/ 40,202 (33.5)	14,144/ 43,057 (32.85)	4734/ 40,202 (11.78)	4934/ 43,057 (11.46)	1014/ 40,202 (2.52)	1050/ 43,057 (2.44)	40,202/ 330,098 (12.18)	43,057/ 330,098 (13.04)
XI ^m	10,999/ 14,892 (73.86)	10,382/ 13,967 (74.33)	3893/ 14,892 (26.14)	3585/ 13,967 (25.67)	1201/ 14,892 (8.06)	1108/ 13,967 (7.93)	217/ 14,892 (1.46)	195/ 13,967 (1.4)	14,892/ 330,098 (4.51)	13,967/ 330,098 (4.23)
XII ⁿ	430/583 (73.76)	252/348 (72.41)	153/583 (26.24)	96/348 (27.59)	38/583 (6.52)	28/348 (8.05)	5/583 (0.86)	3/348 (0.86)	583/330,098 (0.18)	348/330,098 (0.11)
XIII ^o	991/1397 (70.94)	684/960 (71.25)	406/1397 (29.06)	276/960 (28.75)	130/1397 (9.31)	96/960 (10)	28/1397 (2)	18/960 (1.88)	1397/330,098 (0.42)	960/330,098 (0.29)
XIV ^p	7426/ 10,277 (72.26)	7499/ 10,389 (72.18)	2851/ 10,277 (27.74)	2890/ 10,389 (27.82)	924/ 10,277 (8.99)	927/ 10,389 (8.92)	179/ 10,277 (1.74)	174/ 10,389 (1.67)	10,277/ 330,098 (3.11)	10,389/ 330,098 (3.15)
XV ^q	29/35 (82.86)	—	6/35 (17.14)	—	2/35 (5.71)	—	1/35 (2.86)	—	35/330,098 (0.01)	0/330,098 (0)
XVI ^r	46/58 (79.31)	3/3 (100)	12/58 (20.69)	0/3 (0)	5/58 (8.62)	0/3 (0)	0/58 (0)	0/3 (0)	58/330,098 (0.02)	3/330,098 (0)
XVII ^s	357/494 (72.27)	181/246 (73.58)	137/494 (27.73)	65/246 (26.42)	38/494 (7.69)	17/246 (6.91)	10/494 (2.02)	3/246 (1.22)	494/330,098 (0.15)	246/330,098 (0.07)
XVIII ^t	11,072/ 16,269 (68.06)	11,641/ 17,075 (68.18)	5197/ 16,269 (31.94)	5434/ 17,075 (31.82)	1802/ 16,269 (11.08)	1879/17,075 (11)	448/ 16,269 (2.75)	454/ 17,075 (2.66)	16,269/ 330,098 (4.93)	17,075/ 330,098 (5.17)

Chapter	No excess mortality, n/N (%)		Excess mortality, n/N (%)		Severe excess mortality (>4 SDs), n/N (%)		Extreme excess mortality (>6 SDs), n/N (%)		Total, n/N (%)	
	Human	AUTOCOD	Human	AU-TOCOD	Human	AU-TOCOD	Human	AU-TOCOD	Human	AUTOCOD
XIX ^u	0/2 (0)	—	2/2 (100)	—	2/2 (100)	—	0/2 (0)	—	2/330,098 (0)	0/330,098 (0)
XX ^v	10,497/ 14,128 (74.3)	9363/ 12,527 (74.74)	3631/ 14,128 (25.7)	3164/ 12,527 (25.26)	1214/ 14,128 (8.59)	1040/ 12,527 (8.3)	273/ 14,128 (1.93)	239/ 12,527 (1.91)	14,128/ 330,098 (4.28)	12,527/ 330,098 (3.79)
—	—	1/1 (100)	—	0/1 (0)	—	0/1 (0)	—	0/1 (0)	0/330,098 (0)	1/330,098 (0)
Total	236,681/ 330,098 (71.7)	236,681/ 330,098 (71.7)	93,417/ 330,098 (28.3)	93,417/ 330,098 (28.3)	30,110/ 330,098 (9.12)	30,110/ 330,098 (9.12)	6240/ 330,098 (1.89)	6240/ 330,098 (1.89)	330,098/ 330,098 (100)	330,098/ 330,098 (100)

^aPercentage values represent the proportion of death certificates for each period analyzed considering the total of each chapter except for the total column, which gives the proportion of each chapter for all the death certificates.

^bCertain infectious and parasitic diseases.

^cNeoplasms.

^dDiseases of the blood and blood-forming organs and certain disorders involving the immune system.

^eEndocrine, nutritional, and metabolic diseases.

^fMental and behavioral disorders.

^gDiseases of the nervous system.

^hDiseases of the eye and adnexa.

ⁱMissing values.

^jDiseases of the ear and mastoid process.

^kDiseases of the circulatory system.

^lDiseases of the respiratory system.

^mDiseases of the digestive system.

ⁿDiseases of the skin and subcutaneous tissue.

^oDiseases of the musculoskeletal system and connective tissue.

^pDiseases of the genitourinary system.

^qPregnancy, childbirth, and the puerperium.

^rCertain conditions originating in the perinatal period.

^sCongenital malformations, deformations, and chromosomal abnormalities.

^tSymptoms, signs, and abnormal clinical and laboratory findings not elsewhere specified.

^uInjury, poisoning, and certain other consequences of external causes.

^vExternal causes of morbidity and mortality.

Table 2. Description of the study population for the 3 most common chapters (II, IX, and X) for all the periods analyzed (N=330,098)^a.

Block	No excess mortality, n/N (%)		Excess mortality, n/N (%)		Severe excess mortality (>4 SDs), n/N (%)		Extreme excess mortality (>6 SDs), n/N (%)		Total, n/N (%)	
	Human	AUTOCOD	Human	AU-TOCOD	Human	AU-TOCOD	Human	AU-TOCOD	Human	AUTOCOD
C00- C97 ^b	63,379/ 84,031 (75.42)	61,770/ 81,919 (75.4)	20,652/ 84,031 (24.58)	20,149/ 81,919 (24.6)	5942/ 84,031 (7.07)	5695/ 81,919 (6.95)	1141/ 84,031 (1.36)	1102/ 81,919 (1.35)	84,031/ 223,459 (37.6)	81,919/ 223,771 (36.61)
D00- D09 ^c	8/9 (88.89)	— ^d	1/9 (11.11)	—	0/9 (0)	—	0/9 (0)	—	9/223,459 (0)	0/223,771 (0)
D10- D36 ^e	224/311 (72.03)	171/224 (76.34)	87/311 (27.97)	53/224 (23.66)	26/311 (8.36)	14/224 (6.25)	7/311 (2.25)	4/224 (1.79)	311/223,459 (0.14)	224/223,771 (0.1)
D37- D48 ^f	1090/1486 (73.35)	954/1319 (72.33)	396/1486 (26.65)	365/1319 (27.67)	120/1486 (8.08)	110/1319 (8.34)	18/1486 (1.21)	24/1319 (1.82)	1486/223,459 (0.66)	1319/223,771 (0.59)
I05- I09 ^g	376/490 (76.73)	301/408 (73.77)	114/490 (23.27)	107/408 (26.23)	40/490 (8.16)	42/408 (10.29)	3/490 (0.61)	4/408 (0.98)	490/223,459 (0.22)	408/223,771 (0.18)
I10- I15 ^h	5291/7611 (69.52)	6210/8938 (69.48)	2320/7611 (30.48)	2728/8938 (30.52)	810/7611 (10.64)	796/8938 (8.91)	149/7611 (1.96)	149/8938 (1.67)	7611/223,459 (3.41)	8938/223,771 (3.99)
I20- I25 ⁱ	14,858/ 21,153 (70.24)	14,803/ 20,979 (70.56)	6295/ 21,153 (20.76)	6176/ 20,979 (29.44)	2093/ 21,153 (9.89)	1925/ 20,979 (9.18)	471/ 21,153 (2.23)	441/ 20,979 (2.1)	21,153/ 223,459 (9.47)	20,979/ 223,771 (9.38)
I26- I28 ^j	1615/2314 (69.79)	1627/2296 (70.86)	699/2314 (30.21)	669/2296 (29.14)	222/2314 (9.59)	197/2296 (8.58)	43/2314 (1.86)	41/2296 (1.79)	2314/223,459 (1.04)	2296/223,771 (1.03)
I30- I52 ^k	18,232/ 26,016 (70.08)	18,563/ 26,565 (69.88)	7784/ 26,016 (29.92)	8002/ 26,565 (30.12)	2490/ 26,016 (9.57)	2433/ 26,565 (9.16)	524/ 26,016 (2.01)	509/ 26,565 (1.92)	26,016/ 223,459 (11.64)	26,565/ 223,771 (11.87)
I60- I69 ^l	24,836/ 34,595 (71.79)	24,133/ 33,625 (71.77)	9759/ 34,595 (28.21)	9492/ 33,625 (28.23)	3162/ 4,595 (9.14)	2893/ 33,625 (8.6)	621/ 34,595 (1.8)	568/ 33,625 (1.69)	34,595/ 223,459 (15.48)	33,625/ 223,771 (15.03)
I70- I79 ^m	3494/4794 (72.88)	3001/4136 (72.56)	1300/4794 (27.12)	1135/4136 (27.44)	431/4794 (8.99)	351/4136 (8.49)	102/4794 (2.13)	79/4136 (1.91)	4794/223,459 (2.15)	4136/223,771 (1.85)
I80- I89 ⁿ	303/427 (70.96)	203/296 (68.58)	124/427 (29.04)	93/296 (31.42)	36/427 (8.43)	23/296 (7.77)	5/427 (1.17)	2/296 (0.68)	427/223,459 (0.19)	296/223,771 (0.13)
I95- I99 ^o	16/20 (80)	9/9 (100)	4/20 (20)	0/9 (0)	3/20 (15)	0/9 (0)	0/20 (0)	0/9 (0)	20/223,459 (0.01)	9/223,771 (0)
J00- J06 ^p	28/46 (60.87)	14/18 (77.78)	18/46 (39.13)	4/18 (22.22)	7/46 (15.22)	1/18 (5.56)	1/46 (2.17)	0/18 (0)	46/223,459 (0.02)	18/223,771 (0.01)
J09- J18 ^q	11,866/18,191 (65.23)	12,417/18,775 (66.14)	6325/18,191 (34.77)	6358/18,775 (33.86)	2248/18,191 (12.36)	2082/18,775 (11.09)	481/18,191 (2.64)	441/18,775 (2.35)	18,191/223,459 (8.14)	18,775/223,771 (8.39)
J20- J22 ^r	1409/2102 (67.03)	1394/2067 (67.44)	693/2102 (32.97)	673/2067 (32.56)	251/2102 (11.94)	228/2067 (11.03)	61/2102 (2.9)	55/2067 (2.66)	2102/223,459 (0.94)	2067/223,771 (0.92)
J30- J39 ^s	42/53 (79.25)	30/40 (75)	11/53 (20.75)	10/40 (25)	6/53 (11.32)	2/40 (5)	0/53 (0)	0/40 (0)	53/223,459 (0.02)	40/223,771 (0.02)
J40- J47 ^t	5929/8953 (66.22)	6814/10,234 (66.58)	3024/8953 (33.78)	3420/10,234 (33.42)	1070/8953 (11.95)	1113/10,234 (10.88)	232/8953 (2.59)	240/10,234 (2.35)	8953/223,459 (4.01)	10,234/223,771 (4.57)
J60- J70 ^u	1649/2340 (70.47)	1692/2345 (72.15)	691/2340 (29.53)	653/2345 (27.85)	211/2340 (9.02)	180/2345 (7.68)	42/2340 (1.79)	33/2345 (1.41)	2340/223,459 (1.05)	2345/223,771 (1.05)
J80- J84 ^v	1168/1636 (71.39)	1102/1544 (71.37)	468/1636 (28.61)	442/1544 (28.63)	157/1636 (9.6)	131/1544 (8.48)	28/1636 (1.71)	27/1544 (1.75)	1636/223,459 (0.73)	1544/223,771 (0.69)

Block	No excess mortality, n/N (%)		Excess mortality, n/N (%)		Severe excess mortality (>4 SDs), n/N (%)		Extreme excess mortality (>6 SDs), n/N (%)		Total, n/N (%)	
	Human	AUTOCOD	Human	AU-TOCOD	Human	AU-TOCOD	Human	AU-TOCOD	Human	AUTOCOD
J85-J86 ^w	155/215 (72.09)	61/83 (73.49)	60/215 (27.91)	22/83 (26.51)	16/215 (7.44)	2/83 (2.41)	1/215 (0.47)	0/83 (0)	215/223,459 (0.1)	83/223,771 (0.04)
J90-J94 ^x	160/221 (72.4)	182/249 (73.09)	61/221 (27.6)	67/249 (26.91)	22/221 (9.95)	17/249 (6.83)	5/221 (2.26)	3/249 (1.2)	221/223,459 (0.1)	249/223,771 (0.11)
J95-J99 ^y	4330/6445 (67.18)	5207/7702 (67.61)	2115/6445 (32.82)	2495/7702 (32.39)	746/6445 (11.57)	745/7702 (9.67)	163/6445 (2.53)	160/7702 (2.08)	6445/223,459 (2.88)	7702/223,771 (3.44)
Total	160,458/223,459 (71.81)	160,658/223,771 (71.8)	63,001/223,459 (28.19)	63,113/223,771 (28.2)	20,109/223,459 (9)	18,980/223,771 (8.48)	4098/223,459 (1.83)	3882/223,771 (1.73)	223,459/223,459 (100)	223,771/223,771 (100)

^aPercentage values represent the proportion of death certificates for each period analyzed considering the total of each block except for the total column, which gives the proportion of each block for all the death certificates.

^bMalignant neoplasms.

^cIn situ neoplasms.

^dMissing values.

^eBenign neoplasms.

^fNeoplasms of uncertain or unknown behavior.

^gChronic rheumatic heart diseases.

^hHypertensive diseases.

ⁱIschemic heart diseases.

^jPulmonary heart disease and diseases of pulmonary circulation.

^kOther forms of heart disease.

^lCerebrovascular diseases.

^mDiseases of the arteries, arterioles, and capillaries.

ⁿDiseases of the veins, lymphatic vessels, and lymph nodes not elsewhere classified.

^oOther and unspecified disorders of the circulatory system.

^pAcute upper respiratory infections.

^qInfluenza and pneumonia.

^rOther acute lower respiratory infections.

^sOther diseases of the upper respiratory tract.

^tChronic lower respiratory diseases.

^uLung diseases owing to external agents.

^vOther respiratory diseases principally affecting the interstitium.

^wSuppurative and necrotic conditions of the lower respiratory tract.

^xOther diseases of the pleura.

^yOther diseases of the respiratory system.

Results for ICD-10 Chapters

The *caret* package provides the confusion matrix, which evaluates AUTOCOD's performance by calculating some performance metrics. The full performance metrics calculated for AUTOCOD can be found in [Multimedia Appendix 1 \[17-19\]](#).

As presented in Table S2 in [Multimedia Appendix 1 \[17-19\]](#), the specificity in all ICD-10 chapters was >0.97 for periods without excess mortality. The highest values of sensitivity (or recall) were for chapter II—"Neoplasms" (0.95), chapter XVIII—"Symptoms, signs, and abnormal clinical and laboratory findings not elsewhere specified" (0.93), and chapter IX—"Diseases of the circulatory system" (0.91). Considering the PPV (or precision), the highest values were for chapter

XVI—"Certain conditions originating in the perinatal period" (1.00), chapter II—"Neoplasms" (0.98), and chapter IX—"Diseases of the circulatory system" (0.92). The highest F_1 -scores were for chapter II—"Neoplasms" (0.96), chapter IX—"Diseases of the circulatory system" (0.91), and chapter XVIII—"Symptoms, signs, and abnormal clinical and laboratory findings not elsewhere specified" (0.90).

Specificity in all ICD-10 chapters was >0.96 for the excess mortality periods. The highest values of sensitivity (or recall) were for chapter II—"Neoplasms" (0.95), chapter XVIII—"Symptoms, signs, and abnormal clinical and laboratory findings not elsewhere specified" (0.93), and chapter IX—"Diseases of the circulatory system" (0.91). Considering the PPV (or precision), the highest values were for chapter

II—“Neoplasms” (0.97), chapter IX—“Diseases of the circulatory system” (0.91), and chapter XVIII—“Symptoms, signs, and abnormal clinical and laboratory findings not elsewhere specified” (0.88). The highest F_1 -scores were for chapter II—“Neoplasms” (0.96), chapter IX—“Diseases of the circulatory system” (0.91), and chapter XVIII—“Symptoms, signs, and abnormal clinical and laboratory findings not elsewhere specified” (0.90).

Specificity in periods with severe excess mortality (>4 SDs) was >0.96 in all ICD-10 chapters. The highest values of sensitivity (or recall) were for chapter II—“Neoplasms” (0.94), chapter XVIII—“Symptoms, signs, and abnormal clinical and laboratory findings not elsewhere specified” (0.92), and chapter IX—“Diseases of the circulatory system” (0.91). Considering the PPV (or precision), the highest values were for chapter II—“Neoplasms” (0.97), chapter IX—“Diseases of the circulatory system” (0.91), and chapter XVIII—“Symptoms, signs, and abnormal clinical and laboratory findings not elsewhere specified” (0.88). The highest F_1 -scores were for chapter II—“Neoplasms” (0.96), chapter IX—“Diseases of the circulatory system” (0.91), and chapter XVIII—“Symptoms, signs, and abnormal clinical and laboratory findings not elsewhere specified” (0.90).

For periods with extreme excess mortality (>6 SDs), specificity in all ICD-10 chapters was >0.96. The highest values of sensitivity (or recall) were for chapter II—“Neoplasms” (0.95), chapter IX—“Diseases of the circulatory system” (0.91), and chapter XVIII—“Symptoms, signs, and abnormal clinical and laboratory findings not elsewhere specified” (0.90). Considering the PPV (or precision), the highest values were for chapter II—“Neoplasms” (0.96), chapter IX—“Diseases of the circulatory system” (0.90), and chapter XVIII—“Symptoms, signs, and abnormal clinical and laboratory findings not elsewhere specified” (0.88). The highest F_1 -scores were for chapter II—“Neoplasms” (0.96), chapter IX—“Diseases of the circulatory system” (0.91), and chapter XVIII—“Symptoms, signs, and abnormal clinical and laboratory findings not elsewhere specified” (0.89).

Considering the weighted average of all chapters, the results we obtained for the performance metrics of AUTOCOD are presented in Table 3. For sensitivity, PPV, and F_1 -score, there was no difference between periods without excess mortality and those with excess mortality (<0.01). There was a decrease of 0.01 from periods without excess mortality to periods with severe excess mortality (>4 SDs). There was a decrease of 0.04 when comparing the weighted average of periods without excess mortality and periods with extreme excess mortality (>6 SDs).

Table 3. Average performance metrics for different periods for the International Statistical Classification of Diseases and Health-Related Problems, 10th Revision, chapter classification of AUTOCOD.

	Sensitivity (weighted average)	Specificity (weighted average)	Positive predictive value (weighted average)	F_1 -score (weighted average)
No excess mortality	0.88	0.98	0.88	0.88
Excess mortality	0.88	0.98	0.88	0.88
Severe excess mortality (>4 SDs)	0.87	0.98	0.87	0.87
Extreme excess mortality (>6 SDs)	0.85	0.94	0.84	0.84

It is vital to analyze the differences between periods without excess mortality and periods of excess mortality, severe excess mortality, or extreme excess mortality and which chapters perform better.

According to Table 4, the biggest differences in the sensitivity values of AUTOCOD between periods without excess mortality and periods with excess mortality were found in chapter XVI—“Certain conditions originating in the perinatal period” (0.07), chapter XVII—“Congenital malformations, deformations, and chromosomal abnormalities” (0.05), chapter VIII—“Diseases of the ear and mastoid process” (−0.07), and chapter XII—“Diseases of the skin and subcutaneous tissue” (−0.08). For the 3 most common chapters, the differences were 0.00 (chapter II—“Neoplasms”), 0.00 (chapter IX—“Diseases of the circulatory system”), and 0.01 (chapter X—“Diseases of the respiratory system”). Regarding the differences in sensitivity values between periods without excess mortality and periods of severe excess mortality (Z score of >4 SDs), the biggest differences were found in chapter VIII—“Diseases of the ear and mastoid process” (−0.22), chapter XII—“Diseases of the

skin and subcutaneous tissue” (−0.12), chapter XVI—“Certain conditions originating in the perinatal period” (0.07), and chapter XVII—“Congenital malformations, deformations, and chromosomal abnormalities” (0.07). For the 3 most common chapters, the differences were 0.01 (chapter II—“Neoplasms”), 0.01 (chapter IX—“Diseases of the circulatory system”), and 0.00 (chapter X—“Diseases of the respiratory system”). When comparing the difference between the sensitivity values of AUTOCOD for periods without excess mortality and periods of extreme excess mortality (Z score of >6 SDs), the biggest differences were found in chapter XVII—“Congenital malformations, deformations, and chromosomal abnormalities” (0.19), chapter III—“Diseases of the blood and blood-forming organs and certain disorders involving the immune system” (0.17), chapter XIII—“Diseases of the musculoskeletal system and connective tissue” (0.10), and chapter XII—“Diseases of the skin and subcutaneous tissue” (0.08). For the 3 most common chapters, the differences were 0.00 (chapter II—“Neoplasms”), 0.00 (chapter IX—“Diseases of the circulatory system”), and 0.00 (chapter X—“Diseases of the respiratory system”).

Table 4. Comparison among sensitivity values of AUTOCOD depending on the period (without excess mortality and with excess mortality, severe excess mortality, or extreme excess mortality) by chapter of the International Statistical Classification of Diseases and Health-Related Problems, 10th Revision.

Chapter	No excess mortality	Excess mortality	Difference (no excess mortality–excess mortality)	KLD ^a (no excess mortality and excess mortality)	Severe excess mortality (>4 SDs)	Difference (>4 SDs–no excess mortality)	KLD (no excess mortality and >4 SDs)	Extreme excess mortality (>6 SDs)	Difference (>6 SDs–no excess mortality)	KLD (no excess mortality and >6 SDs)
I	0.67	0.67	<0.01	0.00	0.67	0.00	0.00	0.65	0.02	0.02
II	0.95	0.95	0.00	0.00	0.94	0.01	0.01	0.95	0.00	0.00
III	0.57	0.55	0.02	0.02	0.58	0.00	0.00	0.41	0.17	0.19
IV	0.81	0.81	0.00	0.00	0.81	0.00	0.00	0.82	–0.02	–0.02
V	0.77	0.78	0.00	0.00	0.78	0.00	0.00	0.77	0.01	0.01
VI	0.79	0.80	–0.01	–0.01	0.79	0.00	0.00	0.79	0.01	0.01
VII	0.00	0.00	0.00	0.00	N/A ^b	N/A	N/A	N/A	N/A	N/A
VIII	0.18	0.25	–0.07	–0.06	0.40	–0.22	–0.14	N/A	N/A	N/A
IX	0.91	0.91	0.00	0.00	0.91	0.01	0.01	0.91	0.00	0.00
X	0.90	0.89	0.01	0.01	0.90	0.00	0.00	0.89	0.00	0.00
XI	0.80	0.79	0.02	0.02	0.76	0.04	0.04	0.76	0.04	0.04
XII	0.28	0.35	–0.08	–0.07	0.40	–0.12	–0.10	0.20	0.08	0.09
XIII	0.42	0.42	0.00	0.00	0.42	0.00	0.00	0.32	0.10	0.11
XIV	0.76	0.76	0.01	0.01	0.76	0.00	0.00	0.74	0.02	0.02
XV	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
XVI	0.07	0.00	0.07	0.57	0.00	0.07	0.57	N/A	N/A	N/A
XVII	0.39	0.34	0.05	0.06	0.32	0.07	0.08	0.20	0.19	0.26
XVIII	0.93	0.93	0.00	0.00	0.92	0.01	0.01	0.90	0.03	0.03
XIX	N/A	0.00	N/A	N/A	0.00	N/A	N/A	N/A	N/A	N/A
XX	0.79	0.76	0.02	0.02	0.75	0.04	0.04	0.76	0.02	0.02

^aKLD: Kullback-Leibler divergence.

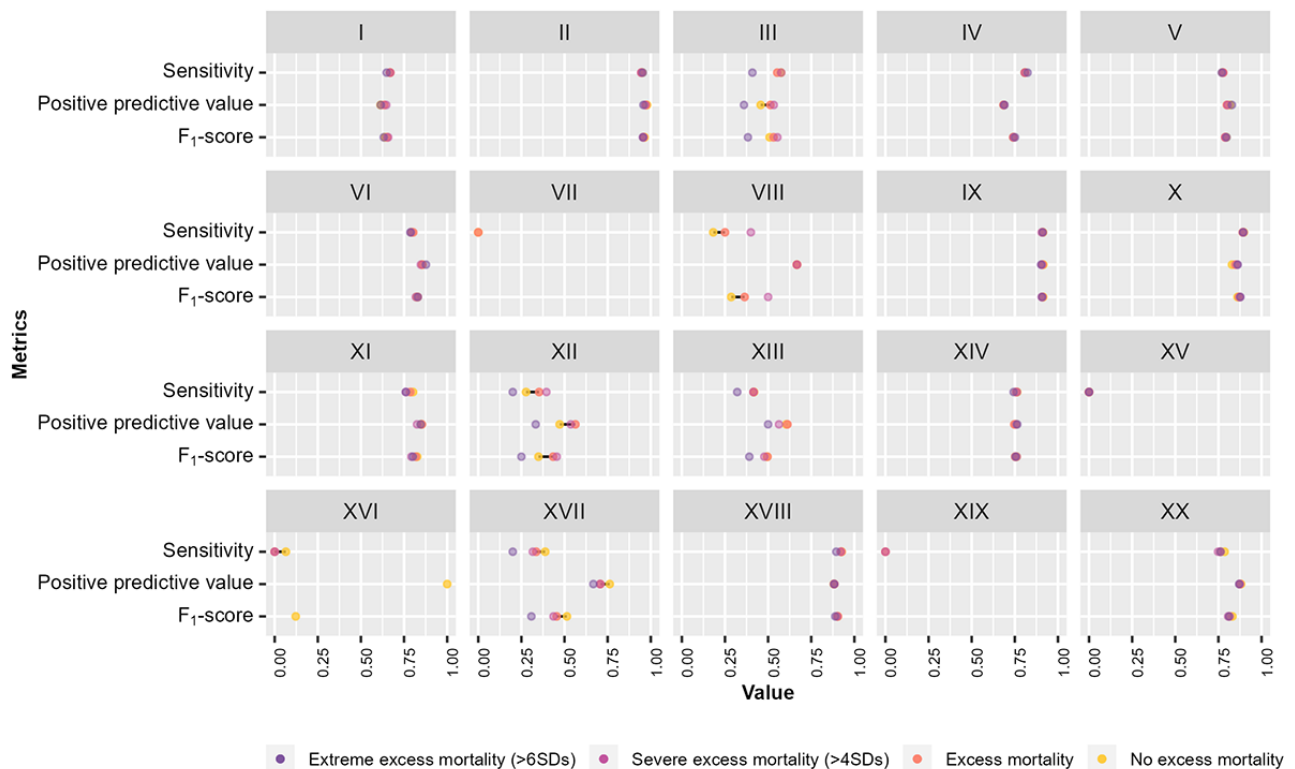
^bN/A: not applicable.

In addition, [Table 4](#) shows the KLD between periods without excess mortality and periods of excess mortality. For 9 chapters, including 2 of the most prevalent (chapter II—“Neoplasms” and chapter IX—“Diseases of the circulatory system”), the KLD was 0, indicating that the distribution of values for periods of excess mortality was similar to that for periods of no excess mortality. For other chapters, such as chapter X—“Diseases of the respiratory system,” the KLD was close to 0. In chapter XVI—“Certain conditions originating in the perinatal period,” the KLD was particularly high, implying a large difference in the probability distributions. Regarding the KLD between periods without excess mortality and periods of extreme excess mortality (Z score of >4 SDs), the sensitivity had a KLD of 0 for 9 chapters, including chapter X—“Diseases of the respiratory

system.” It also had a KLD close to 0 for chapter II—“Neoplasms” and chapter IX—“Diseases of the circulatory system.” When comparing the difference between the KLD for the sensitivity of AUTOCOD for periods without excess mortality and periods of extreme excess mortality (Z score of >6 SDs), sensitivity had a KLD of 0 in the 3 most prevalent chapters as well as chapter XV—“Pregnancy, childbirth, and the puerperium.”

The differences in the performance measures of AUTOCOD between periods without excess mortality and periods of excess or extreme excess mortality are shown in [Figure 2](#). The absolute values of the observations for each period analyzed and additional comparisons of AUTOCOD performance measures can be found in [Multimedia Appendix 1](#) [17-19].

Figure 2. Comparison between performance metrics of AUTOCOD during periods of excess mortality, severe excess mortality, and extreme excess mortality and periods without excess mortality for International Statistical Classification of Diseases and Health-Related Problems, 10th Revision (ICD-10), chapters. DGS: Directorate-General of Health; SICO: Death Certificate Information System.



Last data: August 8, 2019
Source: SICO | Author: DGS

Results for ICD-10 Blocks

This section analyzes the ICD-10 classification by blocks for only the 3 most common chapters (chapter II—“Neoplasms,” chapter IX—“Diseases of the circulatory system,” and chapter X—“Diseases of the respiratory system”).

As presented in Table S3 in [Multimedia Appendix 1 \[17-19\]](#), specificity in all ICD-10 blocks was >0.99 for periods without excess mortality. The highest values of sensitivity (or recall) were for blocks C00-C97—malignant neoplasms (0.98), I60-I69—cerebrovascular diseases (0.94), and J09-J18—influenza and pneumonia (0.94). Considering the PPV (or precision), the highest values were for blocks C00-C97—malignant neoplasms (0.99), I60-I69—cerebrovascular diseases (0.95), and I20-I25—ischemic heart disease (0.94). The highest F_1 -scores were for blocks C00-C97—malignant neoplasms (0.99), I60-I69—cerebrovascular diseases (0.95), and J09-J18—influenza and pneumonia (0.94).

Specificity in all ICD-10 blocks was >0.99 for periods of excess mortality. The highest values of sensitivity (or recall) were for blocks C00-C97—malignant neoplasms (0.98), I60-I69—cerebrovascular diseases (0.94), and J09-J18—influenza and pneumonia (0.93). Considering the PPV (or precision), the highest values were for blocks C00-C97—malignant neoplasms (0.99), I60-I69—cerebrovascular diseases (0.95), and J09-J18—influenza and pneumonia (0.94). The highest F_1 -scores

were for blocks C00-C97—malignant neoplasms (0.98), I60-I69—cerebrovascular diseases (0.95), and J09-J18—influenza and pneumonia (0.94).

Regarding the periods of severe excess mortality, with Z scores of >4 SDs, the specificity in all ICD-10 blocks was >0.98. The highest values of sensitivity (or recall) were for blocks C00-C97—malignant neoplasms (0.97), I60-I69—cerebrovascular diseases (0.93), and J09-J18—influenza and pneumonia (0.93). Considering the PPV (or precision), the highest values were for blocks C00-C97—malignant neoplasms (0.99), I60-I69—cerebrovascular diseases (0.96), and I20-I25—ischemic heart disease (0.95). The highest F_1 -scores were for blocks C00-C97—malignant neoplasms (0.98), I60-I69—cerebrovascular diseases (0.95), and I20-I25—ischemic heart diseases (0.94).

Specificity in all ICD-10 blocks was >0.99 for periods of extreme excess mortality (z score of >6 SDs). The highest values of sensitivity (or recall) were for blocks C00-C97—malignant neoplasms (0.97), I60-I69—cerebrovascular diseases (0.93), and J09-J18—influenza and pneumonia (0.93). Considering the PPV (or precision), the highest values were for blocks D10-D36—benign neoplasms (1.00); I80-I89—diseases of the veins, lymphatic vessels, and lymph nodes not elsewhere classified (1.00); and C00-C97—malignant neoplasms (0.99). The highest F_1 -scores were for blocks C00-C97—malignant neoplasms (0.98), I60-I69—cerebrovascular diseases (0.94), and J09-J18—influenza and pneumonia (0.94).

Table 5 presents AUTOCOD's performance metrics for the weighted average of all the blocks analyzed. For sensitivity, PPV, and F_1 -score, there was a decrease of 0.01 from periods without excess mortality to periods with excess mortality, severe excess mortality (>4 SDs), and extreme excess mortality (>6 SDs).

Considering the differences between periods of excess mortality and periods without excess mortality, it is important to analyze which blocks had the biggest differences.

According to **Table 6**, the largest differences in the sensitivity of AUTOCOD between periods without excess mortality and periods of excess mortality were in block J00-J06—acute upper respiratory infections (0.34), J30-J39—other diseases of the upper respiratory tract (0.28), and I95-I99—other and unspecified disorders of the circulatory system (0.08). Regarding the difference in sensitivity between periods without excess mortality and periods of severe excess mortality (>4 SDs), the largest differences were in block J00-J06—acute upper respiratory infections (0.41), J85-J86—suppurative and necrotic conditions of the lower respiratory tract (0.23), J30-J39—other diseases of the upper respiratory tract (0.20), and I05-I09—chronic rheumatic heart diseases (−0.22). The largest differences in the sensitivity of AUTOCOD between periods without excess mortality and periods of extreme excess mortality (>6 SDs) were in blocks J00-J06—acute upper respiratory infections (0.41), J85-J86—suppurative and necrotic conditions of the lower respiratory tract (0.31), and I05-I09—chronic rheumatic heart diseases (−0.26).

Table 6 also shows the KLD between periods without excess mortality and periods of excess mortality. For 7 blocks,

including C00-C97—malignant neoplasms and I60-I69—cerebrovascular diseases, the KLD was 0. Several blocks had values of KLD very close to 0, such as I20-I25—ischemic heart diseases and J09-J18—influenza and pneumonia. When comparing the difference between the KLD for the sensitivity of AUTOCOD for periods without excess mortality and periods of extreme excess mortality (Z score of >4 SDs), sensitivity had a KLD of 0 in 2 blocks: D37-D48—neoplasms of uncertain or unknown behavior and J95-J99—other diseases of the respiratory system. It also showed a KLD very close to 0 in blocks such as C00-C97—malignant neoplasms and I60-I69—cerebrovascular diseases. Regarding the KLD between periods without excess mortality and periods of extreme excess mortality (Z score of >6 SDs), the sensitivity had a KLD of 0 for I26-I28—pulmonary heart disease and diseases of pulmonary circulation and J40-J47—chronic lower respiratory diseases and a KLD very close to 0 for C00-C97—malignant neoplasm, I20-I25—ischemic heart diseases, and J09-J18—influenza and pneumonia. Some blocks, such as J00-J06—acute upper respiratory infections and J85-J86—suppurative and necrotic conditions of the lower respiratory tract, had a particularly high KLD for increasing mortality periods.

The differences in the performance measures of AUTOCOD among periods without excess mortality, with excess mortality, and with extreme excess mortality according to ICD-10 blocks are shown in **Figure 3**. Additional AUTOCOD performance comparisons between periods can be found in **Multimedia Appendix 1** [17-19].

Table 5. Weighted averages of performance metrics for different periods for the International Statistical Classification of Diseases and Health-Related Problems, 10th Revision, block classification of AUTOCOD.

	Sensitivity (weighted average)	Specificity (weighted average)	Positive predictive value (weighted average)	F_1 -score (weighted average)
No excess mortality	0.94	0.99	0.94	0.94
Excess mortality	0.93	0.99	0.93	0.93
Severe excess mortality (>4 SDs)	0.93	0.99	0.93	0.93
Extreme excess mortality (>6 SDs)	0.93	0.99	0.93	0.93

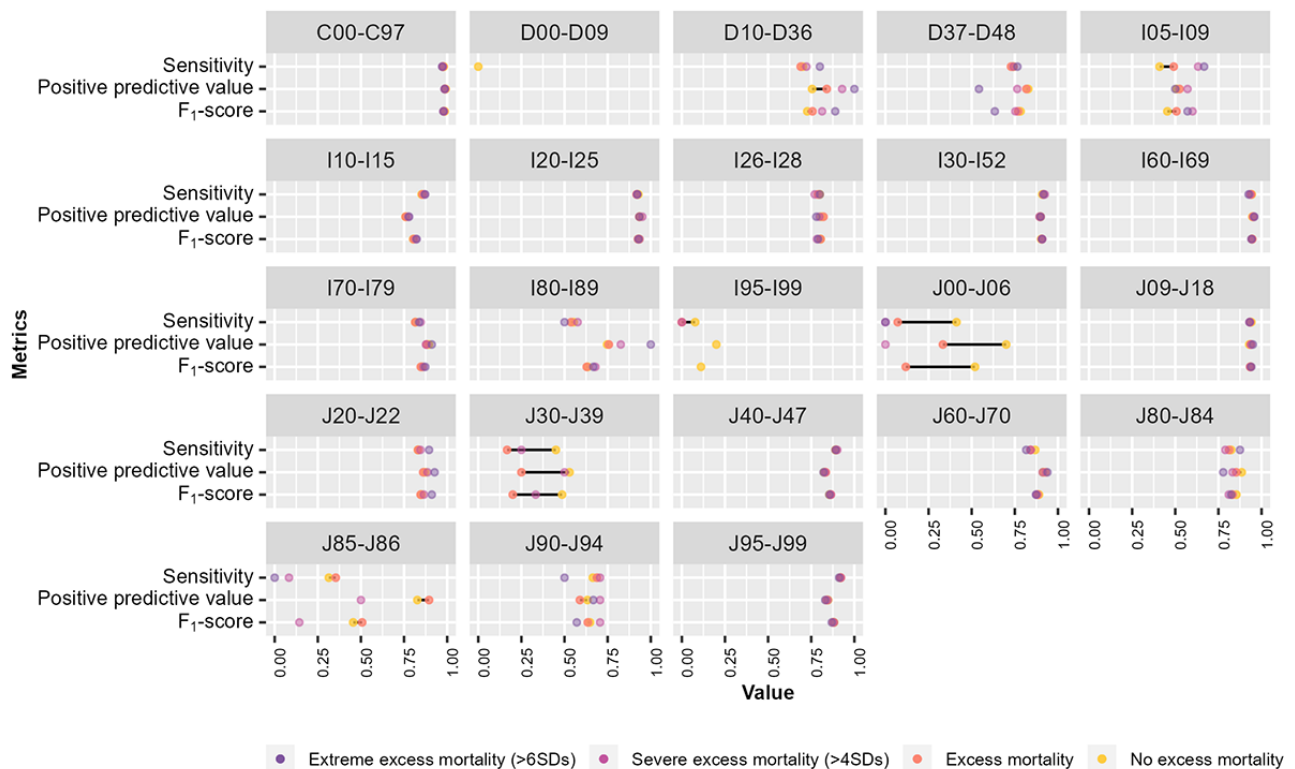
Table 6. Comparison between sensitivity values of AUTOCOD depending on the period (total, excess mortality, or without excess mortality) by International Statistical Classification of Diseases and Health-Related Problems, 10th Revision, block.

Block	No excess mortality	Excess mortality	Difference (no excess mortality–excess mortality)	KLD ^a (no excess mortality and excess mortality)	Severe excess mortality (>4 SDs)	Difference (>4 SDs–no excess mortality)	KLD (no excess mortality and >4 SDs)	Extreme excess mortality (>6 SDs)	Difference (>6 SDs–no excess mortality)	KLD (no excess mortality and >6 SDs)
C00-C97	0.98	0.98	<0.01	0.00	0.97	<0.01	0.01	0.97	<0.01	0.01
D00-D09	0.00	N/A ^b	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
D10-D36	0.70	0.69	0.01	0.01	0.72	-0.02	-0.02	0.80	-0.10	-0.09
D37-D48	0.74	0.73	0.01	0.01	0.74	0.00	0.00	0.77	-0.02	-0.02
I05-I09	0.41	0.49	-0.08	-0.07	0.63	-0.22	-0.18	0.67	-0.26	-0.20
I10-I15	0.85	0.86	-0.01	0.00	0.87	-0.02	-0.02	0.87	-0.02	-0.02
I20-I25	0.93	0.92	0.01	0.01	0.92	0.01	0.01	0.92	0.01	0.01
I26-I28	0.80	0.79	0.01	0.01	0.77	0.03	0.03	0.80	-<0.01	0.00
I30-I52	0.91	0.92	-0.01	-0.01	0.92	-0.02	-0.01	0.92	-0.01	-0.01
I60-I69	0.94	0.94	<0.01	0.00	0.93	0.01	0.01	0.93	0.02	0.02
I70-I79	0.82	0.82	<0.01	0.00	0.85	-0.03	-0.03	0.84	-0.02	-0.02
I80-I89	0.55	0.54	0.02	0.02	0.58	-0.02	-0.02	0.50	0.05	0.06
I95-I99	0.08	0.00	0.08	0.69	0.00	0.08	0.69	N/A	N/A	N/A
J00-J06	0.41	0.07	0.34	0.72	0.00	0.41	4.38	0.00	0.41	4.38
J09-J18	0.94	0.93	0.01	0.01	0.93	0.01	0.01	0.93	0.01	0.01
J20-J22	0.83	0.83	0.00	0.00	0.85	-0.01	-0.01	0.90	-0.06	-0.06
J30-J39	0.45	0.17	0.28	0.45	0.25	0.20	0.26	N/A	N/A	N/A
J40-J47	0.89	0.89	0.00	0.00	0.90	-0.01	-0.01	0.89	0.00	0.00
J60-J70	0.87	0.84	0.03	0.03	0.84	0.03	0.03	0.82	0.05	0.05
J80-J84	0.82	0.81	0.01	0.01	0.79	0.03	0.03	0.88	-0.05	-0.05
J85-J86	0.31	0.35	-0.04	-0.04	0.08	0.23	0.42	0.00	0.31	4.38
J90-J94	0.66	0.69	-0.02	-0.02	0.71	-0.04	-0.04	0.50	0.16	0.19
J95-J99	0.92	0.92	0.00	0.00	0.92	0.00	0.00	0.91	0.01	0.01

^aKLD: Kullback-Leibler divergence.

^bN/A: not applicable.

Figure 3. Comparison between performance metrics of AUTOCOD during periods of excess mortality and periods without excess mortality for International Statistical Classification of Diseases and Health-Related Problems, 10th Revision (ICD-10), blocks. DGS: Directorate-General of Health; SICO: Death Certificate Information System.



Last data: August 8, 2019
Source: SICO | Author: DGS

Discussion

Principal Findings

Continuous and systematic mortality data collection is crucial for monitoring the population's health and complementing epidemiological studies. This national study is the first to demonstrate the robustness of deep neural networks in classifying primary causes of death even during periods of excess mortality, enabling cause-specific mortality surveillance, which is not widely performed worldwide. This study demonstrated a consistently good performance of AUTOCOD in different periods regardless of excess mortality rates. The results demonstrate the potential of AI algorithms to expedite disease classification and coding, making them a valuable tool for real-time surveillance, timely assessment of public health risks, and planification of responses. Proving that these algorithms can operate effectively despite external factors in different environments reinforces the case for their implementation.

AUTOCOD showed high sensitivity (≥ 0.75) in 10 chapters, with values of >0.90 for the 3 most common ones (chapter II—"Neoplasms," chapter IX—"Diseases of the circulatory system," and chapter X—"Diseases of the respiratory system," which together account for 223,459/330,098, 67.69% of all human-codified causes of death). The weighted average of sensitivity in the ICD-10 chapter analysis showed no difference between periods without excess mortality and periods of excess mortality, a difference of 0.01 between periods without excess

mortality and periods of severe excess mortality (>4 SDs), and a difference of 0.04 between periods without excess mortality and periods of extreme excess mortality (>6 SDs). Regarding the ICD-10 block analysis, it showed a difference of 0.01 for the weighted average of sensitivity between periods without excess mortality and periods of excess mortality between periods without excess mortality and periods of severe (at the >4 SD threshold) and between periods without excess mortality and periods of extreme excess mortality (at the >6 SD threshold).

In the different periods considered for the ICD-10 chapter analysis, AUTOCOD showed a consistently good performance, demonstrating a sensitivity (or recall), a PPV (or precision), and an F_1 -score as high as 0.88 for periods without excess mortality and periods of excess mortality and as low as 0.84 in periods of extreme excess mortality (>6 SDs). When we considered only the most common chapters (chapter II—"Neoplasms," chapter IX—"Diseases of the circulatory system," and chapter X—"Diseases of the respiratory system"), sensitivity ranged from 0.94 to 0.95 in chapter II, 0.91 in chapter IX, and 0.89 to 0.90 in chapter X in the different periods analyzed. The same happened with the PPV, which ranged from 0.96 to 0.98 in chapter II, 0.90 to 0.92 in chapter IX, and 0.83 to 0.86 in chapter X. Regarding the F_1 -score, the performance of AUTOCOD was 0.96 in chapter II, 0.91 in chapter IX, and 0.86 to 0.88 in chapter X. When we considered only the most common blocks—C00-C97 (malignant neoplasms), I60-I69 (cerebrovascular diseases), I30-I52 (other forms of heart disease), I20-I25 (ischemic heart diseases), and J09-J18

(influenza and pneumonia)—the sensitivity ranged from 0.91 to 0.98, the PPV ranged from 0.89 to 0.99, and the F_1 -score ranged from 0.90 to 0.99.

AUTOCOD presented high specificity and negative predictive values in all the analyses performed. This was expected as the number of true negatives was consistently much higher than that of true positives. This is not a characteristic of AUTOCOD itself but rather a result of our handling of the sample and our interpretation of the question as a classification problem with a one-versus-all solution. This method is widely used for multiple-output class classification problems. In our case, the individual ICD-10 chapters or blocks were handled as if they were in a binary model, thus assessing each class individually against all the other classes in the model.

It should be noted that chapter XVII (“Symptoms, signs, and abnormal clinical and laboratory findings not elsewhere specified”) consistently presented high performance metrics in AUTOCOD. This does not translate to a correct certification of the cause of death, but it could imply that, when human coders have difficulties classifying the cause of death, so does the AUTOCOD.

These results are aligned with those of previous studies using AUTOCOD [12,13] and, in general, with the literature on deep neural networks applied to the automatic classification of DCs [14,27,28]. Falissard et al [14] developed a deep neural network for automated coding of the underlying cause of death with a test accuracy of 0.978 (95% CI 0.977-0.979) and an F -measure value of 0.952 (95% CI 0.946-0.957) [27]. The proposed approach by Della Mea et al [28] for automated coding of causes of death had an accuracy of 0.990 (95% CI 0.990-0.991) and a macroaveraged accuracy and F_1 -score of 0.974 and 0.968, respectively. Similarly to our study, Della Mea et al [28] found that accuracy was low for chapters with rare causes of death and, therefore, rare causes of death could be ignored.

However, to the best of our knowledge, this is the first time that a deep neural network that classifies basic causes of death has been evaluated while comparing its performance across different time frames according to their excess mortality rates.

Automatic classification of DCs relies on natural language processing (NLP) techniques and algorithms. NLP can translate free text written by the physician who certified the death into classification codes based on the ICD-10. However, this process depends on the text quality of the analyzed DCs. By text quality, we mean how successfully we can automatically classify, retrieve, or extract information from them [29]. Thus, text quality does not involve a single aspect but combines numerous criteria, including spelling, grammar, organization, informative nature, and page layout [30]. Extracting these attributes can become problematic in low-quality texts (poor grammar, many abbreviations, and short sentences). This is a known problem in medical and clinical texts such as patient records or DCs [30]. The performance of systems that rely on attributes of text quality, such as NLP, affects the overall performance of the algorithms—a text of bad quality may result in poor-quality prediction results. To overcome this limitation, after the development AUTOCOD, a processing layer has been added

to the neural network that has the ability to always read words in text fields as the closest word the model knows (eg, for the word *Alzheimer*, it currently identifies >25 ways of misspelling it). Therefore, this processing layer can help minimize text field errors or abbreviations in periods of excess mortality [31-33].

Our results suggest that, even in periods of excess, severe, and extreme excess mortality when the volume of deaths and the pressure on health services might increase, with a consequent impact on physicians that certify deaths and a potential impact on the quality of the text in the DC, AUTOCOD’s performance remains unhindered. It is important to consider analyzing the linguistic properties of the DC, such as variations in text size and the number of fields filled in by physicians, in future work.

Limitations

An important limitation of this study is that the human coders had access to the automatic classification of the DC by AUTOCOD, meaning that the gold standard we used in this research might be biased by the same algorithm we were trying to evaluate. However, this implementation only entered production on July 26, 2019, meaning that manual classification was unbiased for most of the data sets used in this study.

In addition, there is the matter of ICD-10 code ambiguity. This is a known limitation of the ICD-10 for human coders and automatic algorithms of classification that the sometimes discrete differences between codes for similar causes of death can explain. This might explain the difference in sensitivity between, for example, respiratory blocks such as J00-J06 (acute upper respiratory infections) and J09-J18 (influenza and pneumonia), with the latter presenting a less ambiguous cause of death when compared with the former both for human classification and automatic classification. These unspecified codes are not necessarily an error rate but an indicator of the completeness of clinical information of DCs in which sufficient clinical information is not known or available to assign a more specific code. In the case of human coders, it is common that they look for more clinical information in electronic health records. However, AUTOCOD is restricted to the information included in the DC. This stresses the importance of a well-filled and detailed DC by the physician that certifies the death even in periods of excess mortality.

Routinely, racial and ethnic or socioeconomic groups are not collected in the DC. Although other proxies of social vulnerability can be used, such as the municipality of residence, the focus of this research was not the study of differences in subgroups, making this an important next step of investigation.

The human coders that we set as our ground truth were not mistake free. Current research puts the reliability of human coders at approximately 70% to 89% (reliability is a measure for calculating agreement between coders and the consistency of each coder individually) [34]. These performance scores can be in part explained by the use of different codes for similar diseases. Moreover, the DGS has had a range of human coders that varies in number, typically from 4 to 6, and in experience in classifying causes of death. This may also affect the reliability and accuracy of the ground-truth labels we used in this study. Only 1 human coder classifies each DC, and the DGS regularly

conducts an in-house auditing process in which 2 human coders check for internal reliability by classifying a small sample of DCs.

Another possible limitation, known in the field of AI algorithms, is the generalization of our results to other countries [35]. This question of model transferability requires further study. However, we feel confident that our results can be generalized to other algorithms that rely on NLP for automatic classification without a profound impact on the model's performance even in periods of excess mortality.

Strengths

In Portugal, Law 15/2012 of April 3, 2012, established the SICO, a mortality information system based on the electronic registration of DCs [36]. Since then, SICO has become a widespread tool used by physicians nationally. Therefore, it is a well-established source of data and information related to mortality and an international example of the timeliness of mortality statistics [3].

AUTOCOD was built based on the already disseminated existence of DCs in electronic format and has since been validated as an essential tool for the automatic assignment of ICD-10 codes for causes of death [13]. However, this validation never considered differences in periods that might affect the quality of the DC and, consequently, the performance of AUTOCOD. The method we used for evaluating the performance of AUTOCOD during periods of excess mortality, severe excess mortality, and extreme excess mortality is a known method for comparison of the performance of a given index test with a given ground truth or gold standard, making a case for the importance of evaluating algorithms and models in different periods and in the ever-changing environment that might affect the overall performance of the models.

Although the current use of AUTOCOD is limited to supporting human coders, the research findings suggest a compelling case for enhancing the algorithms used for the automated classification of causes of death. In a completed DC, AUTOCOD can be used to accurately classify basic causes of death in real time even in periods of excess mortality, attesting that deep neural networks are robust to eventual changes in the underlying quality of the text. Furthermore, by defining a baseline from the past (and Portugal has digital DC data going back to 2014), we can detect in real time, with high sensitivity, changes in mortality and periods of excess mortality without the need to wait for human classification of cause of death, especially for the more common and less ambiguous causes of death. Finally, with this algorithm, we can use our data to predict

excess deaths that rely on seasonality, such as influenza and pneumonia.

Implications of Our Work

Our work makes a case for using AUTOCOD for real-time mortality surveillance by ICD-10 codes. It can be further validated by other countries wishing to train their neural networks for medical and clinical text classification. Our research also makes a case for auditing, evaluating, and consistently monitoring AI algorithms to identify potential barriers, strengths, and opportunities [37].

As the AUTOCOD algorithm is robust, it can be used to classify the underlying causes of death in periods of excess mortality with no need to wait for manual coding, which allows for adequate real-time cause-specific mortality surveillance, timely assessment of risks to public health, and definition of priorities and planification of responses in both periods with and without excess mortality. This cause-specific mortality surveillance in real time is not carried out widely worldwide and might benefit from further investigation and real-world intervention. This investigation is a step forward in Portugal for the widespread use of the classification of specific causes of death by the AUTOCOD, with renewed confidence in its results regardless of the presence of excess mortality, and for the implementation of targeted public health interventions and practices.

Further investigations should be carried out, such as a comparison of AUTOCOD with other automated coding systems and a new evaluation of the behavior of AUTOCOD during periods of excess mortality caused by the COVID-19 pandemic, including retraining the algorithm with the new codes for COVID-19 that were not present in the ICD-10 when AUTOCOD was built [14,16,28]. To strengthen coding practices, conducting a reliability study among coders at the DGS would also be important.

Conclusions

This study makes the case for deep neural networks as powerful tools for automatically classifying primary causes of death according to the ICD-10 even during periods of excess mortality. Our work could potentially further the use of deep neural networks to facilitate automatic clinical codification, such as of diseases, medical procedures, or DCs. In addition, it may serve as a staple for the real-time monitoring and surveillance of public health threats and problems, allowing for timely action. More broadly, this study highlights the importance of AI algorithms as an advisory tool for public health policies and measures.

Acknowledgments

The authors thank the coders working at the Directorate-General of Health (Isabel Veloso, Liliana Bernardo, Lucília Cardoso, Marina Ramos, and Sofia Pimenta). This study received no specific grants from any funding agency.

Data Availability

Data from the Death Certificate Information System and related analyses are available for research purposes under the conditions foreseen in Law 15/2012.

Conflicts of Interest

None declared.

Multimedia Appendix 1

Description of the data set and AUTOCOD's performance in different periods of excess mortality for both International Statistical Classification of Diseases and Health-Related Problems, 10th Revision, chapters and blocks.

[[PDF File \(Adobe PDF File\), 2631 KB - ai_v2i1e40965_app1.pdf](#)]

References

1. Graphs and maps. EuroMOMO. URL: <https://www.euromomo.eu/graphs-and-maps/> [accessed 2023-04-13]
2. Vestergaard LS, Nielsen J, Richter L, Schmid D, Bustos N, Braeye T, et al. Excess all-cause mortality during the COVID-19 pandemic in Europe - preliminary pooled estimates from the EuroMOMO network, March to April 2020. *Euro Surveill* 2020 Jul;25(26):2001214 [FREE Full text] [doi: [10.2807/1560-7917.ES.2020.25.26.2001214](https://doi.org/10.2807/1560-7917.ES.2020.25.26.2001214)] [Medline: [32643601](https://pubmed.ncbi.nlm.nih.gov/32643601/)]
3. Pinto CS, Anderson RN, Martins H, Marques C, Maia C, Borrvalho MC. Mortality Information System in Portugal: transition to e-death certification. *Eurohealth (Lond)* 2016;22(2):1-53 [FREE Full text] [Medline: [32336930](https://pubmed.ncbi.nlm.nih.gov/32336930/)]
4. SICO - eVM. Vigilância de Mortalidade. URL: https://evm.min-saude.pt/#shiny-tab-info_eVM [accessed 2021-11-08]
5. International Statistical Classification of Diseases and Health-Related Problems, 10th Revision, 5th Edition, 2016. World Health Organization. 2015. URL: <https://apps.who.int/iris/handle/10665/246208> [accessed 2021-11-09]
6. Mazick A, Workshop on mortality monitoring in Europe. Monitoring excess mortality for public health action: potential for a future European network. *Euro Surveill* 2007 Jan 04;12(1):E070104.1. [Medline: [17370927](https://pubmed.ncbi.nlm.nih.gov/17370927/)]
7. Nogueira PJ, Nobre MA, Nicola PJ, Furtado C, Vaz Carneiro A. Excess mortality estimation during the COVID-19 pandemic: preliminary data from Portugal. *Acta Med Port* 2020 Jun 01;33(6):376-383. [doi: [10.20344/amp.13928](https://doi.org/10.20344/amp.13928)] [Medline: [32343650](https://pubmed.ncbi.nlm.nih.gov/32343650/)]
8. Nogueira PJ, Machado A, Rodrigues E, Nunes B, Sousa L, Jacinto M, et al. The new automated daily mortality surveillance system in Portugal. *Euro Surveill* 2010 Apr 01;15(13):19529 [FREE Full text] [Medline: [20394709](https://pubmed.ncbi.nlm.nih.gov/20394709/)]
9. Kanieff M, Rago G, Minelli G, Lamagni T, Sadicova O, Selb J, et al. The potential for a concerted system for the rapid monitoring of excess mortality throughout Europe. *Euro Surveill* 2010 Oct 28;15(43):19697 [FREE Full text] [doi: [10.2807/ese.15.43.19697-en](https://doi.org/10.2807/ese.15.43.19697-en)] [Medline: [21087579](https://pubmed.ncbi.nlm.nih.gov/21087579/)]
10. Hardelid P, Andrews N, Pebody R. Excess mortality monitoring in England and Wales during the influenza A(H1N1) 2009 pandemic. *Epidemiol Infect* 2011 Sep;139(9):1431-1439. [doi: [10.1017/S0950268811000410](https://doi.org/10.1017/S0950268811000410)] [Medline: [21439100](https://pubmed.ncbi.nlm.nih.gov/21439100/)]
11. Simonsen L, Clarke MJ, Stroup DF, Williamson GD, Arden NH, Cox NJ. A method for timely assessment of influenza-associated mortality in the United States. *Epidemiology* 1997 Jul;8(4):390-395. [doi: [10.1097/00001648-199707000-00007](https://doi.org/10.1097/00001648-199707000-00007)] [Medline: [9209852](https://pubmed.ncbi.nlm.nih.gov/9209852/)]
12. Duarte FR. Automated classification of causes of mortality [Thesis]. Instituto Superior Técnico. 2017 Sep. URL: [http://C:/Users/HP/Downloads/Thesis_Automated_Classification_of_Causes_of_Mortality%20\(1\).pdf](http://C:/Users/HP/Downloads/Thesis_Automated_Classification_of_Causes_of_Mortality%20(1).pdf) [accessed 2021-10-26]
13. Duarte F, Martins B, Pinto CS, Silva MJ. Deep neural models for ICD-10 coding of death certificates and autopsy reports in free-text. *J Biomed Inform* 2018 Apr;80:64-77 [FREE Full text] [doi: [10.1016/j.jbi.2018.02.011](https://doi.org/10.1016/j.jbi.2018.02.011)] [Medline: [29496630](https://pubmed.ncbi.nlm.nih.gov/29496630/)]
14. Falissard L, Morgand C, Roussel S, Imbaud C, Ghosn W, Bounebacher K, et al. A deep artificial neural network-based model for prediction of underlying cause of death from death certificates: algorithm development and validation. *JMIR Med Inform* 2020 Apr 28;8(4):e17125 [FREE Full text] [doi: [10.2196/17125](https://doi.org/10.2196/17125)] [Medline: [32343252](https://pubmed.ncbi.nlm.nih.gov/32343252/)]
15. Impact of the Implementation of IRIS software for ICD-10 cause of death coding on mortality statistics, England and Wales. Office of National Statistics. 2014 Aug 8. URL: <https://tinyurl.com/42d6hs9y> [accessed 2021-11-16]
16. Gergonne B, Mazick A, O'Donnell J, Oza A, Cox B, Wuillaume F, et al. A European algorithm for a common monitoring of mortality across Europe. EuroMOMO. URL: https://www.euromomo.eu/uploads/pdf/wp7_report.pdf [accessed 2021-10-11]
17. "Westgard rules" and multirules. Westgard QC. URL: <https://www.westgard.com/mltirule.htm> [accessed 2022-03-21]
18. Kuhn M, Wing J, Weston S, Williams A, Keefer C, Engelhardt A, et al. caret: classification and regression training. The Comprehensive R Archive Network. URL: <https://CRAN.R-project.org/package=caret> [accessed 2022-03-20]
19. Kuhn M. Building predictive models in R using the caret package. *J Stat Softw* 2008;28(5):1-26. [doi: [10.18637/jss.v028.i05](https://doi.org/10.18637/jss.v028.i05)]
20. Nourani V, Sayyah Fard M. Sensitivity analysis of the artificial neural network outputs in simulation of the evaporation process at different climatologic regimes. *Adv Eng Softw* 2012 May;47(1):127-146. [doi: [10.1016/j.advengsoft.2011.12.014](https://doi.org/10.1016/j.advengsoft.2011.12.014)]
21. Drost HG. Philentropy: information theory and distance quantification with R. *J Open Source Softw* 2018 Jun 11;3(26):765. [doi: [10.21105/joss.00765](https://doi.org/10.21105/joss.00765)]
22. R Core Team. R: a language and environment for statistical computing. R Foundation for Statistical Computing. 2017. URL: <https://www.R-project.org/> [accessed 2022-04-05]
23. Wickham H, Chang W, Henry L, Pedersen TL, Takahashi K, Wilke C, et al. ggplot2: create elegant data visualisations using the grammar of graphics. The Comprehensive R Archive Network. URL: <https://CRAN.R-project.org/package=ggplot2> [accessed 2022-04-05]
24. Wickham H, Vaughan D, Girlich M, Ushey K. tidy: tidy messy data. The Comprehensive R Archive Network. URL: <https://CRAN.R-project.org/package=tidy> [accessed 2022-04-05]

25. Wickham H, Averick M, Bryan J, Chang W, McGowan L, François R, et al. Welcome to the Tidyverse. *J Open Source Softw* 2019;4(43):1686. [doi: [10.21105/joss.01686](https://doi.org/10.21105/joss.01686)]
26. Passo 1: check-list de auto-avaliação ética. Escola Nacional de Saúde Pública. 2021. URL: <https://www.ensp.unl.pt/wp-content/uploads/2021/08/passo1-doc1-check-list-de-auto-avaliacao-etica-final-2.pdf> [accessed 2023-04-21]
27. Falissard L, Morgand C, Ghosn W, Imbaud C, Bounebache K, Rey G. Neural translation and automated recognition of ICD-10 medical entities from natural language: model development and performance assessment. *JMIR Med Inform* 2022 Apr 11;10(4):e26353 [FREE Full text] [doi: [10.2196/26353](https://doi.org/10.2196/26353)] [Medline: [35404262](https://pubmed.ncbi.nlm.nih.gov/35404262/)]
28. Della Mea V, Popescu MH, Roitero K. Underlying cause of death identification from death certificates using reverse coding to text and a NLP based deep learning approach. *Inform Med Unlocked* 2020;21:100456. [doi: [10.1016/j.imu.2020.100456](https://doi.org/10.1016/j.imu.2020.100456)]
29. Sonntag D. Assessing the quality of natural language text data. DaimlerChrysler Research and Technology. URL: https://www.dfki.de/~sonntag/text_quality_short.pdf [accessed 2021-11-16]
30. Louis A. Predicting text quality: metrics for content, organization and reader interest. University of Pennsylvania. 2013. URL: <https://repository.upenn.edu/edissertations/665> [accessed 2022-04-18]
31. Wang W, Kreimeyer K, Woo EJ, Ball R, Foster M, Pandey A, et al. A new algorithmic approach for the extraction of temporal associations from clinical narratives with an application to medical product safety surveillance reports. *J Biomed Inform* 2016 Aug;62:78-89 [FREE Full text] [doi: [10.1016/j.jbi.2016.06.006](https://doi.org/10.1016/j.jbi.2016.06.006)] [Medline: [27327528](https://pubmed.ncbi.nlm.nih.gov/27327528/)]
32. Kiefer C. Quality indicators for text data. In: Proceedings of the Workshop on Big (and Small) Data in Science and Humanities. 2019 Presented at: BTW '19; March 4-8, 2019; Rostock, Germany URL: <https://dl.gi.de/server/api/core/bitstreams/89cb2dc4-8a1d-424d-9bce-6569b6e4ae8e/content> [doi: [doi:10.18420/btw2019-ws-15](https://doi.org/10.18420/btw2019-ws-15)]
33. Zhu Y, Sha Y, Wu H, Li M, Hoffman RA, Wang MD. Proposing causal sequence of death by neural machine translation in public health informatics. *IEEE J Biomed Health Inform* 2022 Apr;26(4):1422-1431. [doi: [10.1109/jbhi.2022.3163013](https://doi.org/10.1109/jbhi.2022.3163013)]
34. Harteloh P, de Bruin K, Kardaun J. The reliability of cause-of-death coding in The Netherlands. *Eur J Epidemiol* 2010 Aug;25(8):531-538 [FREE Full text] [doi: [10.1007/s10654-010-9445-5](https://doi.org/10.1007/s10654-010-9445-5)] [Medline: [20309611](https://pubmed.ncbi.nlm.nih.gov/20309611/)]
35. Schwaighofer A, Quinonero-Candela J, Sugiyama M, Lawrence ND. *Dataset Shift in Machine Learning*. New York, NY: Penguin Random House LLC; 2008.
36. Lei n.º 15/2012. Diário da República. URL: <https://dre.pt/web/guest/pesquisa/-/search/554389/details/maximized> [accessed 2021-10-26]
37. Oakden-Rayner L, Gale W, Bonham TA, Lungren MP, Carneiro G, Bradley AP, et al. Validation and algorithmic audit of a deep learning system for the detection of proximal femoral fractures in patients in the emergency department: a diagnostic accuracy study. *Lancet Digit Health* 2022 May;4(5):e351-e358. [doi: [10.1016/s2589-7500\(22\)00004-8](https://doi.org/10.1016/s2589-7500(22)00004-8)]

Abbreviations

AI: artificial intelligence

DC: death certificate

DGS: Directorate-General of Health

EuroMOMO: European mortality monitoring project

ICD-10: International Statistical Classification of Diseases and Health-Related Problems, 10th Revision

KLD: Kullback-Leibler divergence

NLP: natural language processing

PPV: positive predictive value

SICO: Death Certificate Information System

Edited by K El Emam; submitted 11.07.22; peer-reviewed by SJC Soerensen, Z Li, MJ Silva; comments to author 29.01.23; revised version received 21.04.23; accepted 02.06.23; published 22.11.23.

Please cite as:

Pita Ferreira P, Godinho Simões D, Pinto de Carvalho C, Duarte F, Fernandes E, Casaca Carvalho P, Loff JF, Soares AP, Albuquerque MJ, Pinto-Leite P, Peralta-Santos A

Real-Time Classification of Causes of Death Using AI: Sensitivity Analysis

JMIR AI 2023;2:e40965

URL: <https://ai.jmir.org/2023/1/e40965>

doi: [10.2196/40965](https://doi.org/10.2196/40965)

PMID: [38875558](https://pubmed.ncbi.nlm.nih.gov/38875558/)

©Patrícia Pita Ferreira, Diogo Godinho Simões, Constança Pinto de Carvalho, Francisco Duarte, Eugénia Fernandes, Pedro Casaca Carvalho, José Francisco Loff, Ana Paula Soares, Maria João Albuquerque, Pedro Pinto-Leite, André Peralta-Santos.

Originally published in JMIR AI (<https://ai.jmir.org>), 22.11.2023. This is an open-access article distributed under the terms of the Creative Commons Attribution License (<https://creativecommons.org/licenses/by/4.0/>), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR AI, is properly cited. The complete bibliographic information, a link to the original publication on <https://www.ai.jmir.org/>, as well as this copyright and license information must be included.

Original Paper

Determinants of Intravenous Infusion Longevity and Infusion Failure via a Nonlinear Model Analysis of Smart Pump Event Logs: Retrospective Study

Arash Kia^{1*}, PhD; James Waterson^{2*}, BA, MMedEd, MHEc; Norma Bargary^{1*}, PhD; Stuart Rolt^{3*}, BA; Kevin Burke^{1*}, PhD; Jeremy Robertson^{4*}, BSc, BEng; Samuel Garcia^{5*}, BHSc; Alessio Benavoli^{6*}, PhD; David Bergström⁷, PhD

¹Department of Mathematics & Statistics, University of Limerick, Limerick, Ireland

²Medical Affairs, Medication Management Solutions, Becton Dickinson, Dubai, United Arab Emirates

³Medical Affairs, International Infusion Solutions, Becton Dickinson, Winnersh, United Kingdom

⁴Systems Engineering, International Infusion Solutions, Becton Dickinson, Limerick, Ireland

⁵Medical Affairs, Medication Management Solutions, Becton Dickinson, Seville, Spain

⁶School of Computer Science and Statistics, Trinity College Dublin, Dublin, Ireland

⁷Research and Development, Infusion Acute Care, Becton Dickinson, Limerick, Ireland

*these authors contributed equally

Corresponding Author:

James Waterson, BA, MMedEd, MHEc

Medical Affairs

Medication Management Solutions, Becton Dickinson

11F Blue Bay Tower

Business Bay

Dubai, 25229

United Arab Emirates

Phone: 971 566035154

Email: james.waterson@bd.com

Abstract

Background: Infusion failure may have severe consequences for patients receiving critical, short-half-life infusions. Continued interruptions to infusions can lead to subtherapeutic therapy.

Objective: This study aims to identify and rank determinants of the longevity of continuous infusions administered through syringe drivers, using nonlinear predictive models. Additionally, this study aims to evaluate key factors influencing infusion longevity and develop and test a model for predicting the likelihood of achieving successful infusion longevity.

Methods: Data were extracted from the event logs of smart pumps containing information on care profiles, medication types and concentrations, occlusion alarm settings, and the final infusion cessation cause. These data were then used to fit 5 nonlinear models and evaluate the best explanatory model.

Results: Random forest was the best-fit predictor, with an F_1 -score of 80.42, compared to 5 other models (mean F_1 -score 75.06; range 67.48-79.63). When applied to infusion data in an individual syringe driver data set, the predictor model found that the final medication concentration and medication type were of less significance to infusion longevity compared to the rate and care unit. For low-rate infusions, rates ranging from 2 to 2.8 mL/hr performed best for achieving a balance between infusion longevity and fluid load per infusion, with an occlusion versus no-occlusion ratio of 0.553. Rates between 0.8 and 1.2 mL/hr exhibited the poorest performance with a ratio of 1.604. Higher rates, up to 4 mL/hr, performed better in terms of occlusion versus no-occlusion ratios.

Conclusions: This study provides clinicians with insights into the specific types of infusion that warrant more intense observation or proactive management of intravenous access; additionally, it can offer valuable information regarding the average duration of uninterrupted infusions that can be expected in these care areas. Optimizing rate settings to improve infusion longevity for continuous infusions, achieved through compounding to create customized concentrations for individual patients, may be possible in light of the study's outcomes. The study also highlights the potential of machine learning nonlinear models in predicting outcomes and life spans of specific therapies delivered via medical devices.

KEYWORDS

intravenous infusion; vascular access device; alarm fatigue; intensive care units; intensive care; neonatal; predictive model; smart pump; smart device; health device; infusion; intravenous; nonlinear model; medical device; therapy; prediction model; artificial intelligence; AI; machine learning; predict; predictive; prediction; log data; event log

Introduction

Overview

Critical care areas require frequent administration of high-alert, critical, short-half-life infusions, intravenous nutrition, sedation and analgesia, as well as other infusions that require rigorous maintenance for continuous delivery. Outside of the intensive care unit (ICU), approximately 60% of all patients will receive an intravenous infusion during their stay [1].

Abrupt and unexpected infusion failure may have severe consequences for patients if the medications are critical, with short-half-life infusions [2]. Continued interruptions to infusions and infusions not running “to time” can also lead to subtherapeutic management. For example, patients receiving antibiotics who require therapeutic drug monitoring based on metrics like area under the concentration-time curve and trough levels often need blood draws before and after administration. The documented time of administration and subsequent blood draws are commonly based on the prescribed regimen and not on the actual completion of the infusion [3,4]. Infusion downstream and upstream occlusion alarms, when substantial, may also contribute to alarm fatigue among clinical staff [5,6]. One extensive study showed that venous access occlusion alarms are responsible for 55% of all intravenous infusion pump alarms in neonatal ICUs [2].

The issue of infusion failure and its determinants has not been comprehensively studied in the literature. Existing *in vivo* studies have focused on mechanical causes at the vascular access device site [7] and incompatibility issues, either between medications [8] or between medications and administration line materials [9].

A vascular access device (VAD) is defined by the Infusion Nurses Society of the United States as a “catheter, tube, or device inserted into the vascular system, including veins, arteries, and bone marrow” [10]. Definitions for VAD failure include situations where the catheter stops working safely before its intended dwell time or before the traditional 72- to 96-hour dwell time limit [11,12]. Recent guidelines from the Centers for Disease Control and Prevention state that peripheral VADs do not need to be electively resited “more frequently than every 72 to 96 hours” [13]. Using these definitions, the VAD failure rate has been suggested to be as high as 63%, with mean and median values of 46% and 43%, respectively, across studies [14].

The VAD failure rate has a fundamental relationship with the administration method, with gravity administration having a VAD failure rate twice that of even simple rate control infusion devices [15,16]. Modern infusion devices with increased accuracy for the detection of downstream occlusion issues would be expected to reduce the VAD failure rate further. The

management of vascular access and infusions also results in a substantial nursing workload. The Therapeutic Intervention Scoring System-28 allocates 3 points to “multiple intravenous medications” (ie, more than 1 medication, “either as single shots or continuously”), 3 points to any “single vasoactive medication,” 4 points in the case of “multiple vasoactive medications, regardless of types and doses,” and 2 points for the care of a “central venous line.” Therefore, continuous infusions of critical, short-half-life intravenous medications via a central VAD could consume 5 to 9 points from a maximum workload of 46 points that can be undertaken by 1 nurse [17], equating to 10%-19% of a critical care nurse’s total activity time. In a study on nursing workload in ICUs with an average length of stay of 7.7 days, it was found that the mean score based on the Therapeutic Intervention Scoring System-28 was 23 (range 14-32 points) and that nursing time constituted the largest economic cost for ICUs [18].

A 2019 study [19] indicated an “excessive nursing workload” across ICUs that was significantly associated with quality of care. Reducing the number of interventions nurses need to undertake to avoid infusion interruption and to increase infusion longevity would be expected to reduce the baseline of nursing workload in intensive care, high-dependency units, and lower-acuity care areas.

In a 2021 study [20] on the impact of infusion alerts and alarms on nursing workflow, alarms and alerts from both intermittent and continuous infusions were analyzed. Alerts, such as those generated by the Dose Error Reduction System due to dose or rate selection by the clinician outside of the defined limits for individual medications, do not interrupt infusions. The study deemed alerts and alarms as “undesirable error states” and described specific conditions that would interrupt infusions, such as flow occlusion and air-in-line alarms.

In our study, we developed working definitions for infusion longevity, and conversely, infusion failure as follows: infusion longevity may be described as the length of time during which a continuous infusion runs without an alarm state causing unexpected and unplanned interruption to the infusion and comes to an end as a planned cessation. Infusion failure may be described as an infusion that does not reach a planned cessation without clinician interventions to manage unexpected and unplanned interruptions.

Objectives

This study aimed to identify and rank determinants of the longevity of continuous infusions delivered by syringe drivers through the use of nonlinear predictive models to evaluate key factors, and subsequently, develop and test a model for predicting the likelihood of successful infusion longevity; this also involves determining the best predictive model for future use. We expected the analysis to show therapeutic practices and

pump management processes that may assist with infusion longevity. Additionally, we aimed to determine which medications are more likely to cause infusion failure and may warrant more intense observation or access management. We also sought to identify critical care units in which infusion failure is more likely to occur and to assess the likelihood of uninterrupted infusion that can be expected in these care areas.

Methods

Ethical Considerations

We collected infusion data from smart syringe pumps of type CareFusion/BD Alaris Plus CC from different hospitals in Spain. These data are part of a larger data aggregation for the European region, which is held as a repository as part of the obligation for medical device manufacturers to maintain vigilant postmarket surveillance programs for regulatory and quality purposes.

These data are collected passively as part of the standard function of infusion pumps, capturing all events including alerts, alarms, fault conditions, and programming of all infusions. No patient data are recorded. As these data are retrospective, and

therefore, cannot influence clinician decision-making, do not record any direct information related to individual patient therapy, and are detached from any patient or clinician information, there was no requirement for formal ethics approval. The Medication Management Solutions Medical Affairs Department of the pumps' manufacturer gave clearance for using these data in this study. The question of any conflict of interest was also addressed at this stage. None was found, as the variables studied are universal to "smart" infusion pumps and are not exclusive to the pumps studied.

Procedure

The data set was obtained from 384 pumps and contained information about various variables. These variables include the profile, indicating the hospital care unit or ward; medication name or type; infusion rate; medication concentration; syringe brand and syringe size; occlusion setting, indicating the pressure threshold at which the pump alarms for an occlusion; and a configured category label for a dependent variable, indicating if the infusion ended by an unexpected and unplanned occlusion or as a planned cessation. [Table 1](#) shows the values for each categorical variable in our data set.

Table 1. Values for different categorical variables in "Hospitals infusion data set: Spain."

Variable	Skewness statistics
Infusions	158,620
Medications	423
Profiles	42
Syringe brands	11
Occlusions	80,764

We used 5 nonlinear models to fit the data and evaluated them with test data to find the best-fit model. These nonlinear models were the following:

- Random forest: a tree-based ensemble learning method that combines multiple decision trees to make predictions. It has been widely used in medical applications due to its ability to handle complex data sets with high performance [21,22].
- XGBoost: a gradient boosting algorithm that uses a series of weak decision trees so that each tree improves the prediction of the previous one. It is known for its speed and ability to handle large data sets [23].
- K-nearest neighbor (KNN): a nonlinear model that makes predictions based on the closest neighbors to the data point. It is often used for classification and regression problems [24].
- Naive Bayes: a probabilistic algorithm that makes predictions based on Bayes' theorem. It is commonly used for many applications, including medical data sets. The algorithm's naive assumption is that there is independence among input variables of the model [25].
- Support vector machine (SVM): a kernel-based algorithm that separates data points by finding the best hyperplane that maximizes the margin between classes. It is often used for classification and regression problems [26].

Choosing the best machine learning model to be used in a study among the hundreds of different available models should be based on their characteristics and their previous success in the field. We chose 2 different ensemble models with extreme gradient boosting along with the random forest model. These models differ in use, and they allowed us to combine multiple models to reach a result. These are well-known models that can be used as delegates of ensemble learning methods. We used KNN as a delegate for nonparametric instance-based learning models. SVM was used as the most commonly used kernel-based learning model. Naive Bayes was tested to check a Bayesian learning model with an independence assumption between the predictors. This set of models covered a large area of different learning natures, and the ideal model selection was made based on finding a global optimum. Undertaking a trial-and-error procedure among hundreds of models with infinite parameter selection was beyond the scope of our resources, but selecting a starting set of different models that represented different learning algorithms gave us a diverse and comprehensive starting point.

SVM's kernel is a radial basis function with the regularization parameter set to 1. XGBoost uses 100 estimators with both the learning rate and maximum depth set to 1. Our random forest uses 100 decision tree estimators, and it uses the Gini index function as its criterion to measure the split quality in each tree.

The nearest neighbor (K) was set to 3 for the KNN model. The Naive Bayes model used a Gaussian function. The parameters were tested on a validation set of 20% of the entire data set before running the final output of sample testing.

We then evaluated the performance of each model using an F_1 -score and a 5-fold cross validation. F_1 -score is a widely used performance metric in classification tasks that measures the balance between precision and recall. It is the harmonic mean of precision and recall, which means that it takes into account both false positives and false negatives, giving equal weight to both [27]. The F_1 -score ranges from 0 to 1, where a score of 1 represents perfect precision and recall, and a score of 0 represents poor performance.

The formula for the F_1 -score is as follows:



In this formula, precision is the number of true positives divided by the sum of true positives and false positives, and recall is the number of true positives divided by the sum of true positives and false negatives.

The F_1 -score was introduced by Van Rijsbergen [28] in 1979 as a way to evaluate the effectiveness of information retrieval systems; since then, it has been widely adopted in various fields, including natural language processing, machine learning, and

computer vision. F_1 -score is particularly useful when the data set has an imbalance, implying a significant difference in the number of instances for each class; it takes into account both precision and recall, which can be affected by imbalanced data sets.

The target variable was binary, with an imbalance ratio (IR) of 1.05. The IR is defined as the ratio of the majority class to the minority class and is the alternative to skewness in binary classifiers. As a rule of thumb, all IRs less than 1.5 are considered to represent balanced data sets [29,30]. As for the skewness of predictor variables, they can only affect the performance of the models and not the selection of the F_1 -score as an evaluation method, as the F_1 -score is calculated on the target variable and encompasses both precision and recall. Variables such as profile, medication, and syringe brand are categorical, and variables like infusion rate, concentration dose, occlusion setting, and syringe size are continuous. Therefore, we chose to limit ourselves to calculating only the Pearson skewness ratio statistics for continuous (numerical) variables (Table 2). The formula used to calculate the skewness ratio is as follows:

$$\text{Skewness ratio} = (3(\text{mean}(x) - \text{median}(x))) / (\text{standard deviation}(x))$$

The skewness ratios showed that there is no high skewness present in the predictor variables.

Table 2. Skewness statistics with imbalance ratios for the numerical data. On the target variable, the data set had an imbalance ratio of 1.05.

Variable	Skewness statistics		
	Mean	Median	Skewness ratio
Infusion rate	6.28	2.0	1.05
Concentration dose	13.11	2.0	1.22
Syringe size	49.67	50	-0.27
Occlusion setting	175.86	200	-0.27

In this study, we set a default threshold of 0.5 to transform predicted probabilities into binary class labels. This approach, commonly used in similar studies, balances precision and recall. Although not a parameter within the models, this threshold selection is an essential postprocessing step that substantially influences categorizing instances as positive or negative.

The efficacy of our selected 0.5 thresholds is substantiated by the balanced precision and recall rates observed in our results. This aligns effectively with our research objectives. We understand that different applications may require different thresholds, but we suggest that our choice of 0.5 is appropriate due to its consistent performance across various models and data sets. As part of our future endeavors, we are keen to investigate dynamic threshold selections. We recognize that this could significantly influence our study's outcomes.

The best-performing model was chosen as the final analysis model. We also calculated the F_1 -score for a model that consistently resulted in the majority class in the data set, which is the occlusion class in our data set. We called this model the "majority voting model."

In the Results section the selection of random forest as the best-fit model for our data is explained. These results derive from the 5-fold cross-validation technique, where we divided the data set into 5 equal subsets. With each test, 1 subset was run as the test set while we attempted to fit our model with the other 4 subsets as the training set. The 5-fold cross-validation technique is a good modelling practice because it helps to mitigate the problem of overfitting and provides a more accurate estimation of model performance. It is a commonly used approach because it balances the trade-off between the number of folds and the variance in the estimated performance metrics [25].

Once we identified random forest as the best-fit model, we used it to calculate each variable's importance to infusion longevity and to identify the most important predictors of unexpected infusion failure. Variable importance measures the contribution of each variable to the model's overall fitness power. Random forest is a popular machine learning algorithm that combines multiple decision trees to make more accurate predictions. One important aspect of random forest is the calculation of feature

importance, which helps to identify which features have the most impact on the prediction.

Feature importance is calculated by analyzing the contribution of each variable in the decision-making process of each individual tree within the random forest model. The importance of a feature is determined by calculating the total reduction of the impurity measure achieved by splitting on that variable across all trees in the forest [21]. In other words, variables that are able to create the largest reduction in impurity (eg, Gini index or entropy) are considered the most important variables.

The importance scores of each variable are then normalized to ensure that they add up to 1, so they can be compared to one another. This enables researchers to identify which features are most relevant for predicting the target variable or model fitness.

In summary, variable importance in random forest is calculated by measuring the impact of each variable in the decision-making process of each individual tree and then aggregating these values across all trees in the forest. The resulting scores can help researchers to identify the most important features for predicting the target variable [31].

Results

As noted above, random forest was the best-fit model for the data set (Table 3).

As random forest outperformed all other models and had the highest F_1 -score, it was selected to predict infusion occlusion in smart syringe infusion pumps of type CC in “Hospitals infusion data set: Spain.” The results are provided in Figure 1.

Table 3. The F_1 -score of all the selected models’ fits to the infusion data set. The results show that random forest was the best-fit model for our data.

Model	F_1 -score
Majority voting model	67.48
Extreme gradient boosting	79.63
Random forest	80.42
Support vector machine	77.42
Naive Bayes	75.04
K-nearest neighbor	75.73

Figure 1. Variable importance in infusion occlusion prediction for “Hospitals infusion data set: Spain” for CareFusion/BD Alaris Plus syringe pumps.

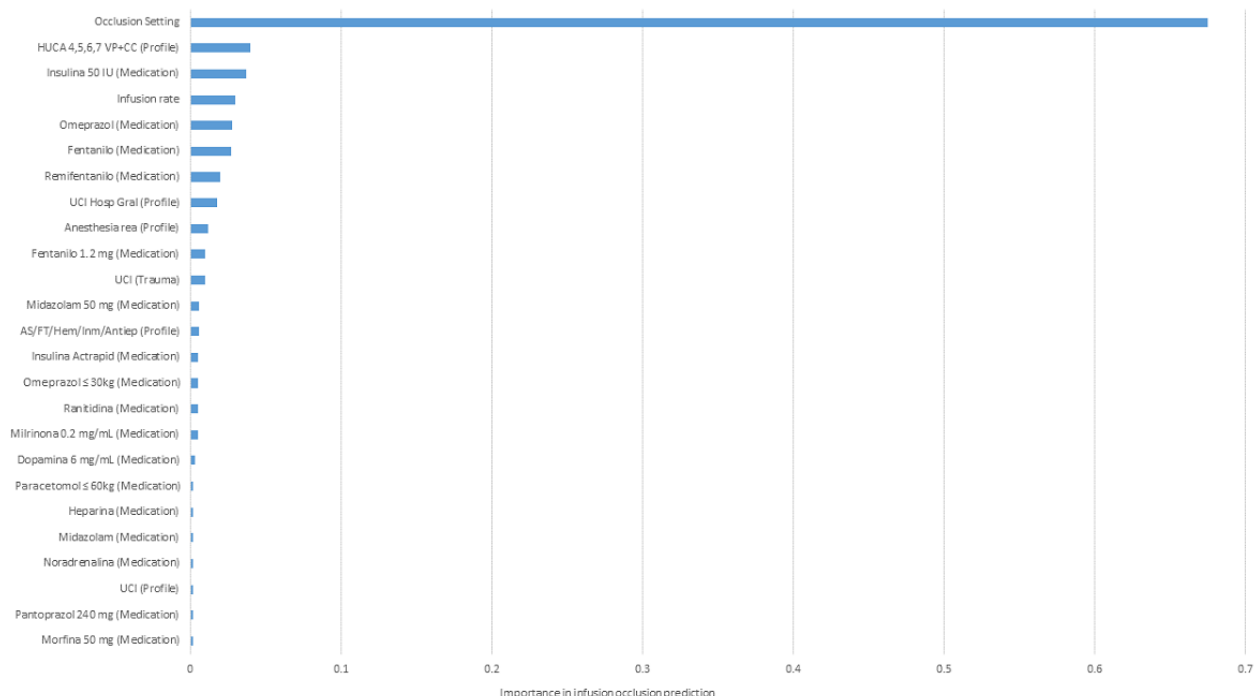


Figure 2A-I shows the total number of infusions with and without occlusions for binary variables, which are essentially bound to the treatment process or location and are beyond the direct control or manipulation of clinicians. The application of profiles differs widely among end-user facilities dependent on their structure, risk strategy, and the services they provide. The

nomenclature is “free text” and is also language dependent. For example, the term “anesthesia rea” seen here would usually pertain to resuscitation area usage and the operating room in several European languages. The profile “HUCA 4.5 6.7 VP+CC” may mean that the hospital has a mix of different pump types from different manufacturers as particular pumps are

mentioned in the profile title. Generally, profiles are given care units, such as neonatal intensive care, adult intensive care, pediatric oncology, and labor and delivery. Therefore, although there is a strong degree of harmonization across facilities and patient types within profiles, as with all multicenter data, there may be differences in acuity; an ICU in a university-level facility will likely have far higher patient acuity levels than a general tertiary care unit. This said, general patient characteristics by profile, in terms of weight, medication concentrations used, and other infusion parameters, may reasonably be expected to be uniform across profiles pertaining to each discipline [2].

Figure 3 illustrates the total number of infusions with and without occlusions across varying values of important nonbinary variables, which are within the control of clinicians or clinical teams.

Figure 4 shows a more detailed breakdown of continuous low-rate infusions by rate. These low-rate infusions are of particular interest and importance clinically, as they are commonly critical short-half-life medication infusions, which are titrated to effect, and their low-rate infusions can cause a longer time to alarm, leading to reduced detectability of “no delivery” states.

Figure 2. Total number of infusions with and without occlusion for binary variables, which are essentially bound to the treatment process or location and beyond the direct control or manipulation of clinicians. (A) HUCA 4.5 6.7 VP+CC (profile). (B) Insulin 1 IU/mL (medication). (C) Omeprazole (medication). (D) Fentanyl (medication). (E) UCI Hosp Gral (profile). (F) Anesthesia Rea (profile). (G) Fentanyl 1.2 Mg (medication). (H) UCI trauma (profile). (I) Remifentanyl (medication).

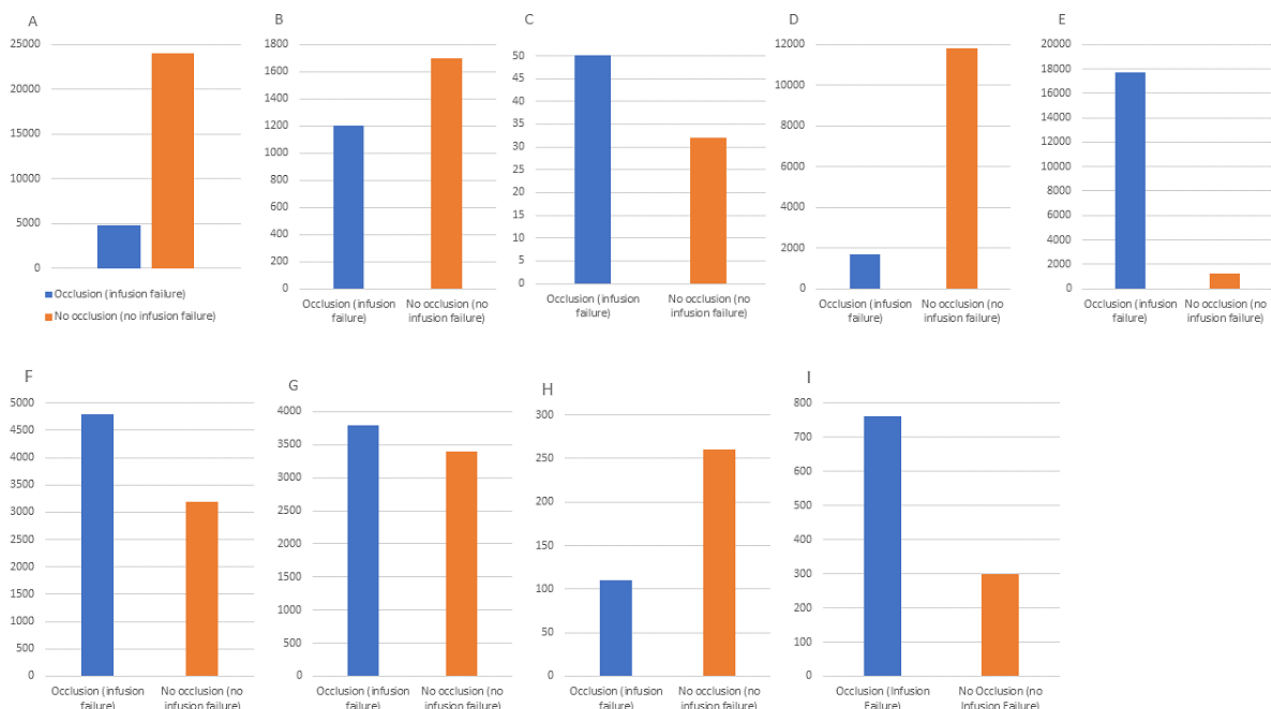


Figure 3. Occlusion versus no occlusion in nonbinary variables of (A) occlusion threshold (mm Hg), (B) infusion rate (mL/hr), and (C) concentration (units/mL). these variables are within the control of clinicians or clinical teams. Concentration units pertain to several units in the International System of Units per ml (eg, mg, mcg, ng, and IU). Blue indicates no occlusion (no infusion failure) and orange indicates occlusion (infusion failure).

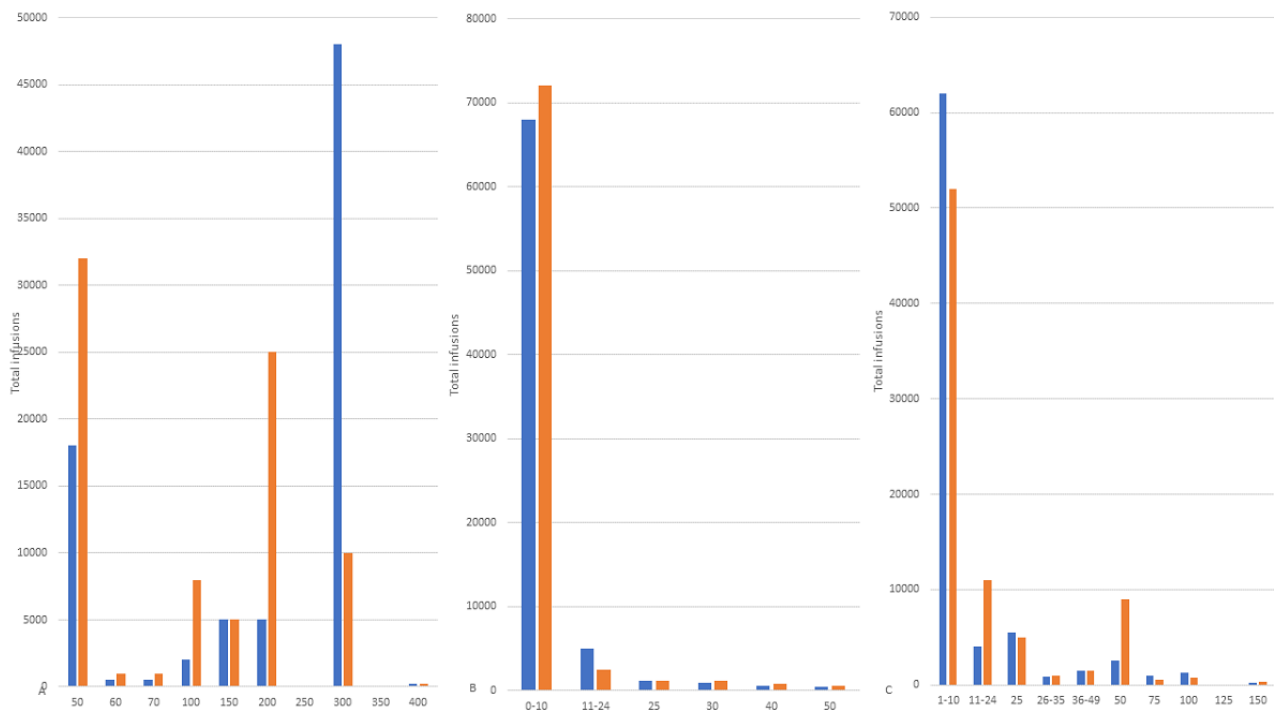
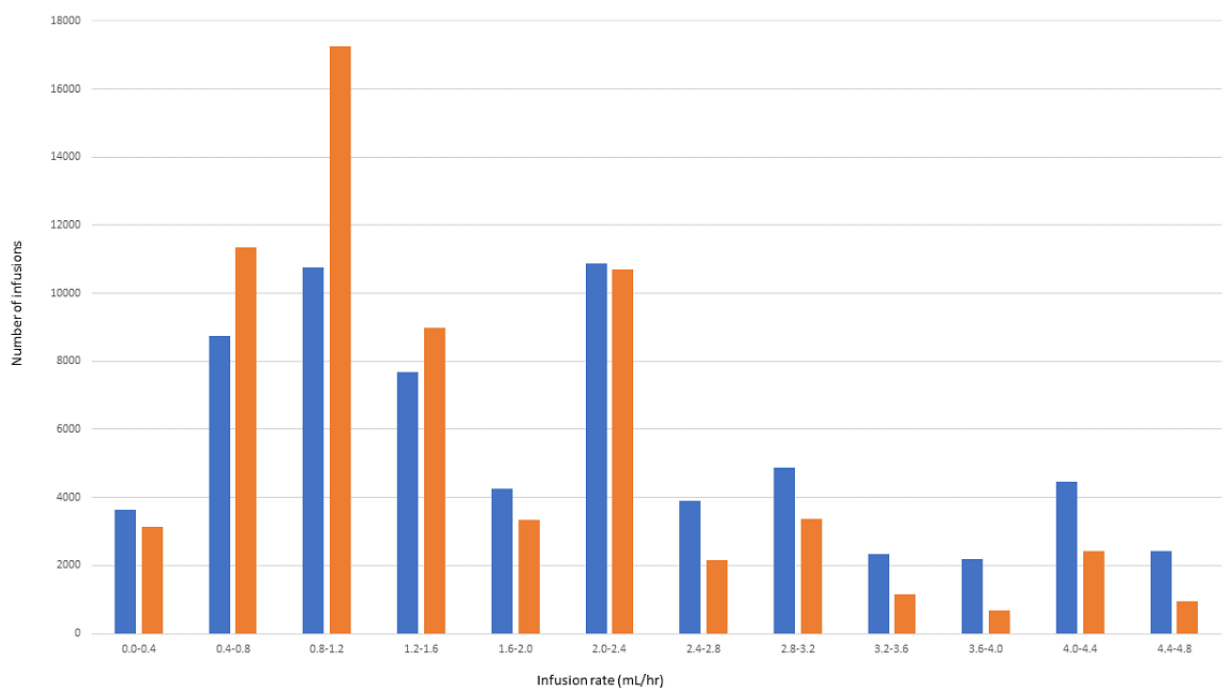


Figure 4. Low-flow infusion rates versus the number of infusions with occlusion (unexpected infusion interruption) and no occlusion (planned infusion cessation). Blue indicates no occlusion and orange indicates occlusion.



Discussion

Study Limitations

Data were only collected from hospitals in Spain, though the investigational method could be applied to other regions as the software deployed in the smart pumps is available worldwide and the structure and deployment of the medication library from

which the data were gathered has been found to be very similar in a previous wide-ranging study of event logs [2].

The study was limited to one type of syringe pump, the CareFusion/BD Alaris CC pump. Similarly, the investigational model can be applied to other pumps, as features like profile, syringe types and brands, medication libraries, and occlusion alarm pressure settings are considered to be universal across

syringe pumps. However, one caveat is that the devices in this study have a fairly unique feature, that of in-line pressure monitoring, where the vein pressure of the patient at the point of the VAD is transmitted directly to a dedicated pressure sensor situated downstream of the syringe and in direct line connection to the VAD. The occlusion alarm setting in this device can be set as low as 15 mm Hg above the detected vein pressure, though commonly, this is set at 30 mm Hg above the vein pressure in neonatal care, and some units use an “auto-offset” feature for automatically setting the alarm after 15 minutes of infusion [32]. Generally, syringe driver infusion pumps without in-line pressure monitoring transmit these data from the drive head behind the syringe, which involves added variables of medication viscosity, syringe friction, and a minimum level of pressure detection, which is rarely below 85 mm Hg.

The question of the type of vascular access of each patient in the study was also a limitation, and whether the VAD was central, peripheral, or umbilical in the case of neonates or a “long line,” such as peripherally inserted central catheters, could not be answered given the data available. However, some inferences could be made from clinically acceptable maximum concentrations for peripheral administration from facility protocols. That being said, evidence is accruing that the concentration of administered individual medications (concentration was not identified as a substantial determinant in our study) is not per se as important as the contact time between medications when multiple infusions pass through a single VAD, along with the subsequent interactions between them. [7,8] At low-flow rates, this contact time is extended, with more time for reactions between medications and subsequent precipitate production. In-line filtration to protect the VAD from precipitate occlusion is emerging from the available evidence as an important factor in determining VAD longevity beyond that of VAD type and medication concentration [7,8].

Including more variables to the information directly available from smart pumps in the analysis, such as direct information about the VAD type and other infusions running via one VAD, may provide a more comprehensive understanding of the in vivo factors that influence infusion longevity.

Principal Findings

The occlusion alarm setting threshold was the most important variable for infusion longevity, and beyond 2 individual medications, the infusion rate was the next most important variable. The data drill-down (Figure 3) and ratio ranking (Tables 4 and 5) show that lower ratios of occlusion to no occlusion were associated with higher rates of infusion, with the rate bracket of >3.6–≤4.0 mL/hr having the best ratio at 0.311. However, the bulk of infusions in the study run at rates far lower than that, with 60.11% running at below 2.0 mL/hr. This is understandable given the fluid balance (or more correctly, the fluid restriction requirements) of medication infusions in critical care, particularly in neonatology and pediatrics. Due to maintaining critical care patients’ nutrition as well as managing renal failure and fluid balance, it is not uncommon to use higher concentration infusions to deliver continuous infusion doses with smaller volumes.

Considering the findings of this study and in vitro studies of infusion startup delay and infusion “no-flow” interruptions [33], as well as the influence of administration line compliance [34], filters [35], the interplay between multiple infusions [36], and resistance from backcheck and antisiphon valves [35], the risks of protracted and clinically important nondelivery and occlusion are likely at low rates, particularly below rates of 0.5 mL/hr [35]. The study’s findings suggest a balance between the need to restrict fluid delivery to patients and maintaining the integrity and longevity of infusion might be best achieved with a rate ranging from >2.0 to ≤2.4 mL/hr (ratio 0.985), although the next higher rate of >2.4–≤2.8 mL/hr would yield a far better ratio of 0.553, albeit with some compromise in fluid restriction control.

Table 4. Ratios of occlusion versus no-occlusion infusions at investigated flow rates (N=131,654).

Rate (mL/hr)	0.0- ≤0.4	>0.4- ≤0.8	>0.8- ≤1.2	>1.2- ≤1.6	>1.6- ≤2.0	>2.0- ≤2.4	>2.4- ≤2.8	>2.8- ≤3.2	>3.2- ≤3.6	>3.6- ≤4.0	>4.0- ≤4.4	>4.4- ≤4.8
Ratio (occlusion vs no occlusion)	0.859	1.297	1.604	1.169	0.786	0.985	0.553	0.691	0.488	0.311	0.547	0.387
Occlusion, n (%)	3635 (53.8)	8741 (43.5)	10,759 (38.4)	7686 (46.1)	4266 (56.0)	10,868 (50.4)	3909 (64.4)	4887 (59.1)	2340 (67.2)	2189 (76.3)	4452 (64.6)	2423 (72.1)
No occlusion, n (%)	3121 (46.2)	11,338 (56.5)	17,261 (61.6)	8986 (53.9)	3352 (44.0)	10,708 (49.6)	2160 (35.6)	3376 (40.9)	1142 (32.8)	680 (23.7)	2437 (35.4)	938 (27.9)
Total infusions studied, n (%)	6756 (5.13)	20,079 (15.25)	28,020 (21.28)	16,672 (12.66)	7618 (5.79)	21,576 (16.39)	6069 (4.61)	8263 (6.28)	3482 (2.64)	2869 (2.18)	6889 (5.23)	3361 (2.55)

Table 5. Ratios of occlusion versus no-occlusion infusions, ranked by the best-performing rate according to the ratio (N=131,654). The possible optimal rate ranges are 2.4-2.8 mL/hr (0.553) and 2.0-2.4 mL/hr (0.985).

Ranking by ratio	1	2	3	4	5	6	7	8	9	10	11	12
Rate (mL/hr)	>3.6- ≤4.0	>4.4- ≤4.8	>3.2- ≤3.6	>4.0- ≤4.4	>2.4- ≤2.8	>2.8- ≤3.2	>1.6- ≤2.0	0.0- ≤0.4	>2.0- ≤2.4	>1.2- ≤1.6	>0.4- ≤0.8	>0.8- ≤1.2
Ratio (occlusion vs no occlusion)	0.311	0.387	0.488	0.547	0.553	0.691	0.786	0.859	0.985	1.169	1.297	1.604
Total infusions studied, n (%)	2869 (2.18)	3361 (2.55)	3482 (2.64)	6889 (5.23)	6069 (4.61)	8263 (6.28)	7618 (5.79)	6756 (5.13)	21,576 (16.39)	16,672 (12.66)	20,079 (15.25)	28,020 (21.28)

Therefore, we suggest that, when feasible, some relaxation of fluid restriction and medication concentrations be considered to deliver infusions at rates between 2.0 and 2.8 mL/hr and higher, if at all possible, for continuous infusions. The improvement in the occlusion ratio may be related to the simple volume of medication moving through the VAD and “flushing” it more effectively than very low-rate infusions can achieve, or it could be attributed to the previously mentioned concept of reduced contact time between medications being administered through a single VAD at higher rates. It is possible to target this flow rate range even through wider titration ranges by manipulation of the final concentration of medications. The suggested rate range would also assist with the clinical detectability of nondelivery [33,35].

In Figure 2A-I, binary variables linked to the treatment process or unit type are identified. These variables are essentially beyond the direct control or manipulation of clinicians. However, this information remains valuable as a “high-risk” indicator for individual medications that may benefit from concentration

manipulation to facilitate higher delivery rates, closer observation of the infusion, or central VAD delivery and exclusive-line administration rather than peripheral administration along with multiple infusions. A multidisciplinary approach to the management of such high-risk medications is advocated.

Conclusions

These findings have important implications for health care professionals who use smart infusion pumps to deliver medications to patients. The study may assist health care professionals to make informed decisions regarding the medication to be administered, concentrations to be used, and infusion duration or rate, to improve infusion longevity, reduce the risk of unplanned infusion interruption, and mitigate risks to the VAD.

The study also highlights the potential of machine learning nonlinear models to predict infusion occlusions in smart infusion pumps. The process of selecting the most appropriate model could be applied to studies involving other medical devices.

Acknowledgments

This publication has emanated from research conducted with the financial support of Science Foundation Ireland (grant 16/RC/3918).

Conflicts of Interest

None declared.

References

- Chaplin S. National Patient Safety Agency report: patient safety in the NHS. *Prescriber* 2006 Apr 05;17(7):52-54. [doi: [10.1002/psb.361](https://doi.org/10.1002/psb.361)]
- Waterson J, Bedner A. Types and frequency of infusion pump alarms and infusion-interruption to infusion-recovery times for critical short half-life infusions: retrospective data analysis. *JMIR Hum Factors* 2019 Aug 12;6(3):e14123 [FREE Full text] [doi: [10.2196/14123](https://doi.org/10.2196/14123)] [Medline: [31407667](https://pubmed.ncbi.nlm.nih.gov/31407667/)]
- Krukavskiy A, Franklin E, Bonk C, Howe J, Dixit R, Adams K, et al. Identifying safety hazards associated with intravenous vancomycin through the analysis of patient safety event reports. *PatientSaf* 2020 Mar 17:31-47. [doi: [10.33940/data/2020.3.3](https://doi.org/10.33940/data/2020.3.3)]
- Peyko V, Friedman-Jakubovics M. Novel approach to vancomycin level monitoring: impact of a multidisciplinary monitoring system on timing of vancomycin levels. *Am J Health Syst Pharm* 2018 Feb 01;75(3):121-126. [doi: [10.2146/ajhp160760](https://doi.org/10.2146/ajhp160760)] [Medline: [29371192](https://pubmed.ncbi.nlm.nih.gov/29371192/)]
- Cho O, Kim H, Lee Y, Cho I. Clinical alarms in intensive care units: perceived obstacles of alarm management and alarm fatigue in nurses. *Healthc Inform Res* 2016 Jan;22(1):46-53 [FREE Full text] [doi: [10.4258/hir.2016.22.1.46](https://doi.org/10.4258/hir.2016.22.1.46)] [Medline: [26893950](https://pubmed.ncbi.nlm.nih.gov/26893950/)]
- Matocha D. Reducing infusion pump alarms through structured interventions. *J Assoc Vasc Access* 2018 Jun 2018;23(2):87-95. [doi: [10.1016/j.java.2018.03.002](https://doi.org/10.1016/j.java.2018.03.002)]

7. Shimoyama S, Takahashi D, Arai S, Asami Y, Nakajima K, Ikeda K, et al. A large amount of microscopic precipitates are inevitably injected during infusion therapy without an in-line filter. *Oxf Med Case Reports* 2022 Feb;2022(2):omab134 [FREE Full text] [doi: [10.1093/omcr/omab134](https://doi.org/10.1093/omcr/omab134)] [Medline: [35198221](https://pubmed.ncbi.nlm.nih.gov/35198221/)]
8. Fonzo-Christe C, Bochaton N, Kiener A, Rimensberger P, Bonnabry P. Incidence and causes of infusion alarms in a neonatal and pediatric intensive care unit: a prospective pilot study. *J Pediatr Pharmacol Ther* 2020;25(6):500-506 [FREE Full text] [doi: [10.5863/1551-6776-25.6.500](https://doi.org/10.5863/1551-6776-25.6.500)] [Medline: [32839653](https://pubmed.ncbi.nlm.nih.gov/32839653/)]
9. Al-Jaber R, Samuda N, Chaker A, Waterson J. Critical care nurses' knowledge of correct line types for administration of common intravenous medications: assessment and intervention study. *JMIR Form Res* 2022 Apr 26;6(4):e36710 [FREE Full text] [doi: [10.2196/36710](https://doi.org/10.2196/36710)] [Medline: [35471247](https://pubmed.ncbi.nlm.nih.gov/35471247/)]
10. Gorski L, Hadaway L, Hagle M, Broadhurst D, Clare S, Kleidon T, et al. Infusion therapy standards of practice, 8th edition. *J Infus Nurs* 2021;44(1S Suppl 1):S1-S224. [doi: [10.1097/NAN.0000000000000396](https://doi.org/10.1097/NAN.0000000000000396)] [Medline: [33394637](https://pubmed.ncbi.nlm.nih.gov/33394637/)]
11. Rickard C, Webster J, Wallis M, Marsh N, McGrail M, French V, et al. Routine versus clinically indicated replacement of peripheral intravenous catheters: a randomised controlled equivalence trial. *The Lancet* 2012 Sep;380(9847):1066-1074. [doi: [10.1016/s0140-6736\(12\)61082-4](https://doi.org/10.1016/s0140-6736(12)61082-4)]
12. Webster J, Clarke S, Paterson D, Hutton A, van Dyk S, Gale C, et al. Routine care of peripheral intravenous catheters versus clinically indicated replacement: randomised controlled trial. *BMJ* 2008 Jul 08;337(7662):a339 [FREE Full text] [doi: [10.1136/bmj.a339](https://doi.org/10.1136/bmj.a339)] [Medline: [18614482](https://pubmed.ncbi.nlm.nih.gov/18614482/)]
13. O'Grady N, Alexander M, Dellinger E, Gerberding J, Heard S, Maki D, et al. Guidelines for the prevention of intravascular catheter-related infections. Centers for Disease Control and Prevention. *MMWR Recomm Rep* 2002 Aug 09;51(RR-10):1-29 [FREE Full text] [Medline: [12233868](https://pubmed.ncbi.nlm.nih.gov/12233868/)]
14. Helm R, Klausner J, Klemperer J, Flint L, Huang E. Accepted but unacceptable: peripheral IV catheter failure. *J Infus Nurs* 2019;42(3):151-164. [doi: [10.1097/NAN.0000000000000326](https://doi.org/10.1097/NAN.0000000000000326)] [Medline: [30985565](https://pubmed.ncbi.nlm.nih.gov/30985565/)]
15. Bivins B, Rapp R, Powers P, Butler J, Haack D. Electronic flow control and roller clamp control in intravenous therapy: a clinical comparison. *Arch Surg* 1980 Jan;115(1):70-72. [doi: [10.1001/archsurg.1980.01380010058011](https://doi.org/10.1001/archsurg.1980.01380010058011)] [Medline: [7350888](https://pubmed.ncbi.nlm.nih.gov/7350888/)]
16. Zazzo J, Millat B, Larrieu H. [Peripheral intravenous infusions. Reduction of morbidity owing to an electronic controller of infusion flow rate]. *Presse Med* 1984 Feb 18;13(7):427-428. [Medline: [6230611](https://pubmed.ncbi.nlm.nih.gov/6230611/)]
17. Miranda D, de Rijk A, Schaufeli W. Simplified Therapeutic Intervention Scoring System: the TISS-28 items--results from a multicenter study. *Crit Care Med* 1996 Jan;24(1):64-73. [doi: [10.1097/00003246-199601000-00012](https://doi.org/10.1097/00003246-199601000-00012)] [Medline: [8565541](https://pubmed.ncbi.nlm.nih.gov/8565541/)]
18. Padilha K, Sousa R, Kimura M, Miyadahira A, da Cruz D, Vattimo M, et al. Nursing workload in intensive care units: a study using the Therapeutic Intervention Scoring System-28 (TISS-28). *Intensive Crit Care Nurs* 2007 Jun;23(3):162-169. [doi: [10.1016/j.iccn.2006.07.004](https://doi.org/10.1016/j.iccn.2006.07.004)] [Medline: [17329107](https://pubmed.ncbi.nlm.nih.gov/17329107/)]
19. Chang L, Yu H, Chao Y. The relationship between nursing workload, quality of care, and nursing payment in intensive care units. *J Nurs Res* 2019 Feb;27(1):1-9 [FREE Full text] [doi: [10.1097/jnr.0000000000000265](https://doi.org/10.1097/jnr.0000000000000265)] [Medline: [29613879](https://pubmed.ncbi.nlm.nih.gov/29613879/)]
20. Yu D, Obuseh M, DeLaurentis P. Quantifying the impact of infusion alerts and alarms on nursing workflows: a retrospective analysis. *Appl Clin Inform* 2021 May;12(3):528-538 [FREE Full text] [doi: [10.1055/s-0041-1730031](https://doi.org/10.1055/s-0041-1730031)] [Medline: [34192773](https://pubmed.ncbi.nlm.nih.gov/34192773/)]
21. Breiman L. Random forests. *Machine Learning* 2021;45:5-32 [FREE Full text] [doi: [10.1007/978-1-4899-7687-1_695](https://doi.org/10.1007/978-1-4899-7687-1_695)]
22. Cutler D, Edwards T, Beard K, Cutler A, Hess K, Gibson J, et al. Random forests for classification in ecology. *Ecology* 2007 Nov;88(11):2783-2792. [doi: [10.1890/07-0539.1](https://doi.org/10.1890/07-0539.1)] [Medline: [18051647](https://pubmed.ncbi.nlm.nih.gov/18051647/)]
23. Chen T, Guestrin C. Xgboost: a scalable tree boosting system. In: *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. 2016 Presented at: KDD '16; August 13 - 17; San Francisco, CA p. 785-794. [doi: [10.1145/2939672.2939785](https://doi.org/10.1145/2939672.2939785)]
24. Cover T, Hart P. Nearest neighbor pattern classification. *IEEE Trans Inform Theory* 1967 Jan;13(1):21-27. [doi: [10.1109/TIT.1967.1053964](https://doi.org/10.1109/TIT.1967.1053964)]
25. McCallum A, Nigam K. A comparison of event models for naive bayes text classification. URL: <http://www.kamalnigam.com/papers/multinomial-aaaiws98.pdf> [accessed 2023-08-17]
26. Cortes C, Vapnik V. Support-vector networks. *Mach Learn* 1995 Sep;20(3):273-297. [doi: [10.1007/BF00994018](https://doi.org/10.1007/BF00994018)]
27. Taha AA, Hanbury A. Metrics for evaluating 3D medical image segmentation: analysis, selection, and tool. *BMC Med Imaging* 2015 Aug 12;15(1):29 [FREE Full text] [doi: [10.1186/s12880-015-0068-x](https://doi.org/10.1186/s12880-015-0068-x)] [Medline: [26263899](https://pubmed.ncbi.nlm.nih.gov/26263899/)]
28. Van Rijsbergen R. Information Retrieval. 1979. URL: http://openlib.org/home/krichel/courses/lis618/readings/rijsbergen79_infor_retriev.pdf [accessed 2023-08-17]
29. Fernández A, García S, del Jesus M, Herrera F. A study of the behaviour of linguistic fuzzy rule based classification systems in the framework of imbalanced data-sets. *Fuzzy Sets and Systems* 2008 Sep;159(18):2378-2398. [doi: [10.1016/j.fss.2007.12.023](https://doi.org/10.1016/j.fss.2007.12.023)]
30. Varma S, Simon R. Bias in error estimation when using cross-validation for model selection. *BMC Bioinformatics* 2006 Feb 23;7(1):91 [FREE Full text] [doi: [10.1186/1471-2105-7-91](https://doi.org/10.1186/1471-2105-7-91)] [Medline: [16504092](https://pubmed.ncbi.nlm.nih.gov/16504092/)]
31. Hastie T, Tibshirani R, Friedman J. *The Elements of Statistical Learning Data Mining, Inference, and Prediction*, Second Edition. Springer Link. 2009. URL: <https://link.springer.com/book/10.1007/978-0-387-84858-7> [accessed 2023-08-17]

32. Bergon-Sendin E, Perez-Grande C, Lora-Pablos D, Moral-Pumarega M, Melgar-Bonis A, Peña-Peloché C, et al. Smart pumps and random safety audits in a Neonatal Intensive Care Unit: a new challenge for patient safety. *BMC Pediatr* 2015 Dec 11;15(1):206 [FREE Full text] [doi: [10.1186/s12887-015-0521-6](https://doi.org/10.1186/s12887-015-0521-6)] [Medline: [26654316](https://pubmed.ncbi.nlm.nih.gov/26654316/)]
33. Baeckert M, Batliner M, Grass B, Buehler P, Daners M, Meboldt M, et al. Performance of modern syringe infusion pump assemblies at low infusion rates in the perioperative setting. *Br J Anaesth* 2020 Feb;124(2):173-182 [FREE Full text] [doi: [10.1016/j.bja.2019.10.007](https://doi.org/10.1016/j.bja.2019.10.007)] [Medline: [31864721](https://pubmed.ncbi.nlm.nih.gov/31864721/)]
34. Weiss M, Neff T, Gerber A, Fischer J. Impact of infusion line compliance on syringe pump performance. *Paediatr Anaesth* 2000;10(6):595-599. [doi: [10.1111/j.1460-9592.2000.566ab.x](https://doi.org/10.1111/j.1460-9592.2000.566ab.x)] [Medline: [11186873](https://pubmed.ncbi.nlm.nih.gov/11186873/)]
35. van der Eijk A, van Rens R, Dankelman J, Smit B. A literature review on flow-rate variability in neonatal IV therapy. *Paediatr Anaesth* 2013 Jan;23(1):9-21. [doi: [10.1111/pan.12039](https://doi.org/10.1111/pan.12039)] [Medline: [23057436](https://pubmed.ncbi.nlm.nih.gov/23057436/)]
36. Cassano-Piché A, Fan M, Sabovitch S, Masino C, Easty A, Health Technology Safety Research Team, Institute for Safe Medication Practices Canada. Multiple intravenous infusions phase 1b: practice and training scan. *Ont Health Technol Assess Ser* 2012;12(16):1-132 [FREE Full text] [Medline: [23074426](https://pubmed.ncbi.nlm.nih.gov/23074426/)]

Abbreviations

ICU: intensive care unit

IR: imbalance ratio

KNN: k-nearest neighbor

SVM: support vector machine

VAD: vascular access device

Edited by G Eysenbach, K El Emam; submitted 01.05.23; peer-reviewed by S Mehdipour, M Obuseh, J Ahn; comments to author 10.06.23; revised version received 06.07.23; accepted 21.07.23; published 13.09.23.

Please cite as:

Kia A, Waterson J, Bargary N, Rolt S, Burke K, Robertson J, Garcia S, Benavoli A, Bergström D

Determinants of Intravenous Infusion Longevity and Infusion Failure via a Nonlinear Model Analysis of Smart Pump Event Logs: Retrospective Study

JMIR AI 2023;2:e48628

URL: <https://ai.jmir.org/2023/1/e48628>

doi: [10.2196/48628](https://doi.org/10.2196/48628)

PMID:

©Arash Kia, James Waterson, Norma Bargary, Stuart Rolt, Kevin Burke, Jeremy Robertson, Samuel Garcia, Alessio Benavoli, David Bergström. Originally published in *JMIR AI* (<https://ai.jmir.org>), 13.09.2023. This is an open-access article distributed under the terms of the Creative Commons Attribution License (<https://creativecommons.org/licenses/by/4.0/>), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in *JMIR AI*, is properly cited. The complete bibliographic information, a link to the original publication on <https://www.ai.jmir.org/>, as well as this copyright and license information must be included.

Original Paper

Predicting Patient Mortality for Earlier Palliative Care Identification in Medicare Advantage Plans: Features of a Machine Learning Model

Anne Bowers¹, PhD; Chelsea Drake¹, MA; Alexi E Makarkin¹, MA; Robert Monzyk¹, MA; Biswajit Maity¹, BA; Andrew Telle¹, BBA

Evernorth Health, Inc, St. Louis, MO, United States

Corresponding Author:

Anne Bowers, PhD
Evernorth Health, Inc
One Express Way
St. Louis, MO, 63121
United States
Phone: 1 860 810 6523
Email: anne.bowers@evernorth.com

Abstract

Background: Machine learning (ML) can offer greater precision and sensitivity in predicting Medicare patient end of life and potential need for palliative services compared to provider recommendations alone. However, earlier ML research on older community dwelling Medicare beneficiaries has provided insufficient exploration of key model feature impacts and the role of the social determinants of health.

Objective: This study describes the development of a binary classification ML model predicting 1-year mortality among Medicare Advantage plan members aged ≥ 65 years ($N=318,774$) and further examines the top features of the predictive model.

Methods: A light gradient-boosted trees model configuration was selected based on 5-fold cross-validation. The model was trained with 80% of cases ($n=255,020$) using randomized feature generation periods, with 20% ($n=63,754$) reserved as a holdout for validation. The final algorithm used 907 feature inputs extracted primarily from claims and administrative data capturing patient diagnoses, service utilization, demographics, and census tract-based social determinants index measures.

Results: The total sample had an actual mortality prevalence of 3.9% in the 2018 outcome period. The final model correctly predicted 44.2% of patient expirations among the top 1% of highest risk members (AUC=0.84; 95% CI 0.83-0.85) versus 24.0% predicted by the model iteration using only age, gender, and select high-risk utilization features (AUC=0.74; 95% CI 0.73-0.74). The most important algorithm features included patient demographics, diagnoses, pharmacy utilization, mean costs, and certain social determinants of health.

Conclusions: The final ML model better predicts Medicare Advantage member end of life using a variety of routinely collected data and supports earlier patient identification for palliative care.

(JMIR AI 2023;2:e42253) doi:[10.2196/42253](https://doi.org/10.2196/42253)

KEYWORDS

palliative; palliative care; machine learning; social determinants; Medicare Advantage; Medicare; predict; algorithm; mortality; older adult

Introduction

Background

Approximately 43% of all Medicare beneficiaries are enrolled in Medicare Advantage plans, totaling 24.4 million Americans as of July 2020 [1]. As the Medicare Advantage population lives longer with more chronic conditions, the need for palliative

services and serious illness care management becomes increasingly important [2]. Palliative services in Medicare Advantage refer to (nonhospice) primary, specialty, and supportive care services for individuals with serious advanced illness and complex chronic conditions that are typically delivered in the patient's home or in a clinical outpatient setting. Palliative care not only may provide patients a better quality of life but also can reduce costs by enabling avoidance of

unnecessary hospitalizations, diagnostic and treatment interventions, and intensive and emergency department care [3-6].

Although the need for and engagement with palliative care among older adults and Medicare beneficiaries is growing, these valuable services are often underutilized [7-9]. One major cause of lower uptake involves unreliability in provider identification of patients who are appropriate for palliative care. Research shows a clinician's intuition alone is not the most effective method for recognizing individuals in general practice who could benefit from palliative services [10-12]. Standardized screening tools that rely primarily on diagnostic criteria, medical record information, and patient-reported needs can promote better reliability in clinician identification of palliative patients [13-20]. However, providers and health plans are increasingly leveraging powerful, data-driven machine learning (ML) techniques to help recognize potential candidates for palliative care earlier and more objectively.

Machine Learning for Palliative Care Identification in Medicare

ML is being adopted across hospital and community-based health care settings as a mechanism to guide early identification of older adults in need of palliative services. ML algorithms attain superior predictive performance from using one or more sources of big data for model training, such as routinely collected medical service claims, electronic medical records, and clinical assessment outcomes [21]. The likelihood of patient mortality within a certain time frame is commonly used as the predictive outcome for ML models intending to identify potential palliative service candidates, because patients who are approaching the end of life are most likely to need and benefit from palliative care [22]. Using ML to identify patients for palliative care not only saves clinicians valuable time but may also improve the efficiency of service delivery to those at highest risk. Early models such as the Charleston Comorbidities Index and Elixhauser score incorporated claims and administrative data to predict mortality of hospitalized older patients [23,24]. Since then, ML models trained using big data from claims and electronic medical records of Medicare beneficiaries (aged ≥ 65 years) in nonhospital settings have achieved greater predictive performance, with the area under the receiver operating characteristic curve (AUC) values ranging between 0.79 and 0.97 [25-28]. The predictive power of ML for the early identification of palliative care in nonhospitalized Medicare patients can surpass that of clinical screening tools developed for similar purposes [14,16].

Previous research on ML mortality models for earlier palliative care identification in the Medicare population has mainly focused on optimizing and comparing the performance of different model configurations [6,25-29]. That said, evaluating critical features of ML mortality models is also necessary to understand performance variation among different model configurations relative to the patient population, health care setting, and type of data analyzed. Failing to report on the important feature inputs gives inadequate transparency about how the algorithm reached its stated outcomes based on the sources of training data [30]. ML model feature impact reporting

appears to be more common in studies analyzing hospitalized Medicare patients [31-33] but has been largely neglected in ML studies that focus on nonhospitalized Medicare beneficiaries [25-28]. Moreover, such prior studies have tapped into various data sources including medical claims, electronic medical records, patient demographics, and clinical assessment information for model training and validation [6,25-29]. The extent to which other, nonmedicalized data are incorporated into these ML mortality models remains unclear, in part due to the lack of discussion around feature impacts. For example, social determinants (eg, socioeconomic status, environmental conditions) are known to influence the mortality and health outcomes of older adults [34,35]. However, previous ML studies in the Medicare population do not clearly indicate if nonmedical data, like measures of the social determinants of health (SDOH), were incorporated as algorithm features [6,25-29,31-33,36].

The important individual features of ML mortality models used to identify palliative care need among nonhospitalized older Medicare patients remain underreported in the current research [25-28]. In an aim to fill this knowledge gap, this study describes the important feature outcomes and performance of a ML algorithm that was developed and validated to predict 1-year mortality of older US adults (aged ≥ 65 years) enrolled in Medicare Advantage plans. Our predictive binary classification model was routinely supplied with data extracted from medical claims as well as electronic health records (EHRs), patient demographic information, and location-specific index measures of SDOH for purposes of identifying Medicare Advantage plan members who may need to connect to palliative resources. Through this study, we investigated the following objectives:

- To what extent is the performance of a baseline ML model (demographics-based with high-risk indicators) predicting 1-year mortality of Medicare Advantage plan members (aged ≥ 65 years) improved by adding features capturing patient service utilization, diagnoses, and SDOH?
- What individual features are of top importance in the final ML model iteration?

Methods

Model Development

An ML algorithm predicting 1-year mortality among Medicare Advantage plan members was developed by the team at Cigna, a large US commercial health benefits company. The aim was to create a prognostic ML model of mortality risk that could enhance the process of identifying patients for palliative care, with the long-term goal of increasing engagement with community-based, nonhospice palliative services among adults (aged ≥ 65 years) in Medicare Advantage plans for whom it would be appropriate. Increasing utilization of palliative services can reduce unnecessary high-cost hospital care and improve patient quality of life. An overview of the health plan's process for identifying and connecting with potential palliative care patients is outlined in [Multimedia Appendix 1](#).

The retrospective data used in the analysis were internally sourced from Cigna's proprietary administrative records and claims database. These standard data elements are routinely

collected to fulfill the operational purposes of the health benefits company; claims and administrative data were only extracted for the purposes of developing the ML algorithm post facto. Security measures for personal health information require all data be completely de-identified by a separate internal team prior to any secondary data analysis to protect member confidentiality. Due to the sensitivity and proprietary nature of the information, data cannot be shared externally.

Ethical Considerations

Our study methods were in accordance with the ethical guidelines of the 1975 Declaration of Helsinki, and our reporting conforms to the Guidelines for Developing and Reporting Machine Learning Predictive Models in Biomedical Research [37]. The data used in the analysis were retrospective, deidentified, and not originally collected for research nor model development purposes; data were only extracted to develop the ML algorithm after the fact. An internal ethics committee approved and regularly reviewed the project protocol throughout the model development process.

Sample Inclusion Criteria

Medicare Advantage plan members eligible for inclusion in analysis were all those with continuous health benefits coverage enrollment as of July 1, 2016, through the feature generation period of December 31, 2017, who also had at least one inpatient or outpatient service encounter in their randomly assigned feature generation time frame. Additionally, to be included in the analyzed sample, during the outcomes period (January 1, 2018, through December 31, 2018), patients must have either (1) had continuous enrollment for the 2018 calendar year or (2) became deceased during 2018. This requirement ensured any beneficiaries who disenrolled from their Medicare Advantage plan in 2018 but were not deceased were not counted as patient expirations.

Machine Learning Method and Training Protocol

Various binary classification ML models were considered. Performance was compared using 5-fold cross-validation. A light gradient-boosted tree model (LightGBM) performed best and was selected based on cross-validation log loss (or cross-entropy loss). The protocol analyzed data from a total sample of 318,774 Medicare Advantage plan members. Features were generated using a training cohort (255,020/318,774, 80% of the sample) with a randomized outcomes time period. Models were further applied to a holdout data set (63,754/318,774, 20% of the sample) to validate and assess generalization to new cases. Data were computed using an instance of DataRobot v6.1.2 (Python 3, custom lightgbm model) running on an on-premise Red Hat Enterprise Linux 7.9 (Maipo) server and with variable resources dedicated via Docker containers (4-8 CPUs each with 32-64 GB RAM).

Target Outcome

The model's predicted outcome was defined as any member who expired between January 1, 2018, and December 31, 2018 (1 year). Patients were determined to be deceased based on corresponding plan enrollment data and validation through reporting to the Centers for Medicare and Medicaid Services [38].

Data Sources and Feature Generation

Feature Generation

A SQL script aggregated data to generate predictive features. To determine the date range for model input generation, a randomized cutoff date was assigned to negative and positive cases. We randomized the actual feature generation dates used per customer, so the distribution of start dates was the same for deceased and alive customers. The random date ensured the ML process did not suffer from seasonality and selection bias. Features were built from the 1-year look-back period (ending December 31, 2017) and included 907 unique inputs based on routinely collected data. Data used in model development were information sourced from claims, EHRs, and administrative member records.

Claims

Data from claims were primarily used to generate features representing patient service utilization. Diagnosis information was also extracted from claims. Types of claims data included medical service claims, pharmacy claims, and laboratory encounters. Laboratory encounters were based on medical claims for lab-related Current Procedural Terminology (CPT) codes. The actual clinical outcomes (results) of laboratory tests are not part of claims data and were thus not incorporated into the model.

Electronic Health Records

Medical data were extracted from EHRs to supplement claims in generating 5 features of high-risk service utilization used in the first iteration of the model (ie, occurrence counts of electrocardiograms, kidney disease, sepsis, ventilator usage, and surgeries). Data from EHRs are aggregated through a third-party vendor partner and are used by the health plan for internal care management and care coordination activities. Not all patients had EHR data on record.

Administrative Member Records

Demographic data, as well as information used to calculate measures of SDOH, were extracted from internal administrative member records. Demographic features were patient age (continuous, in years) and gender (male/female). Social determinants index (SDI) scores are a suite of measures in the administrative member record that were developed for internal use. SDI scores are composite measures representing 6 domains of the SDOH: economy, education, language, health, infrastructure, and food access. SDI scores are determined by the member's census tract, which corresponds to the member's residential address and zip code [39]. The data associated with the measures in each domain are sourced from public use data such as the US Census and US Department of Agriculture (see [Multimedia Appendix 2](#)). Total overall weighted and unweighted SDI scores were also included as features in the model.

Data Preprocessing

Sample members must have had at least one countable service utilization claim in the randomized feature generation period. No feature observations were removed due to missing data. The data had some categorical fields, such as gender or a categorical indicator of utilization status, but most features were continuous

and numeric. Numeric data were not transformed (apart from missing value imputation). Most instances of missing numeric data indicated an individual did not experience a particular type of claim, diagnosis, or event (not due to data quality); such instances were manually coded as 0 to avoid missing values and to represent the patient did not experience the event. Beyond this, DataRobot handles the missing value imputation strategy automatically based on the specified type of imputation algorithm. For the selected model configuration (LightGBM), both continuous/numeric and categorical data had imputed values to represent “missing” data. The final model used ordinal encoding for categorical variables that included a separate category for “missing.” The most common type of missing data was SDI scores, which occurred for 4.9% (15,655/318,774) of the sample population. Age (541/318,774) and gender (647/318,774) data were each missing for 0.2% of the sample.

Model Training and Validation

Data were split 80/20 into training and holdout partitions, respectively. Within the training partition, additional subdivisions were made to tune parameters and apply early stopping. In a LightGBM tree-based algorithm, early stopping refers to stopping the training process if the model performance does not improve after some consecutive iterations. First, the training data were split (training split 1) to keep 90% for train and 10% for test; this set was used for early stopping. Next, the data were split yet again to create training split 2; using only the training portion of training split 1, we assigned 70% for training and 30% for testing. Training split 2 was used to tune model parameters (ie, num_leaves). After these parameters were tuned, we returned to training split 1 to tune the number of estimators (n_estimators) using early stopping (early_stopping). Key parameters included learning_rate (0.05), n_estimators (550), num_leaves (16), max_depth (no limit), min_child_samples (10), and early_stopping_rounds (200). Both the training and holdout partitions had similar mortality rates of 4% in 2018, indicating the mortality outcome was not biased nor skewed in either the training or validation step.

Evaluation Measures

Model performance was assessed using AUC, positive predictive value, negative predictive value, true positive rate, true negative rate, average precision, and lift charts focusing on true positives in the top 10% of predictions for the holdout cohort. Based on the data, DataRobot software selected a threshold of 0.16 for comparing the performance metric matrices of the different model iterations. We performed 1-tailed and 2-tailed z tests to evaluate significant differences between model iterations with the addition of features. Model performance outcomes for the training data set (255,020/318,774, 80% of the sample) are located in [Multimedia Appendix 3](#). Performance outcomes for the holdout data set (63,754/318,774, 20% of the sample) are presented herein to validate the model and assess generalization to new cases. We report the ranked order importance and absolute (unnormalized) importance values of the top 20 model input features based on Shapley Additive Explanations (SHAP) values [30,40].

Results

Of the 318,774 patients included in the total sample, 96.1% (306,227/318,774) were determined to be alive, and 3.9% (12,547/318,774) were determined to be deceased during the 2018 outcomes period (see [Table 1](#)). Compared with alive patients, deceased patients were older, had higher rates of chronic health conditions (cancer, dementia, stroke, heart failure, and chronic respiratory disease), and had greater average service utilization including emergency room, pharmacy, and laboratory encounters. Deceased patients also had lower SDI scores on average (weighted and unweighted) compared with alive patients.

[Table 2](#) summarizes the ML model development and performance outcomes for the holdout cohort (63,754/318,774, 20% of the sample). The baseline model, Model 1 (M1), included 2 demographic features (age and gender) and 5 features capturing elements of high-risk utilization. Model 1 achieved an AUC value of 0.736 (95% CI 0.728-0.744), which was significantly better than mortality prediction based on random chance alone ($z=56.4$, $P<.001$). In the next stage of development, Model 2 (M2) was created by adding 894 more input features using service claims that captured patient clinical diagnoses as well as individual medical, laboratory, and pharmacy utilization. The M2 iteration had an AUC value of 0.834 (95% CI 0.828-0.840), which was a significant performance improvement compared with M1 ($z=19.1$, $P<.001$). Model 3 (M3), the final model, added 8 features representing SDOH (SDI scores). M3 had the best performance of all the model iterations, with an AUC value of 0.839 (95% CI 0.833-0.845), showing significant improvement over that of M1 ($z=20.2$, $P<.001$). The final model (M3) also has a high degree of specificity in that it accurately predicted patients who were not deceased (negative predictive value=0.971), with the model's average precision improving with each iteration (from 0.12 to 0.24). Adding the SDI score features to the final model (M3) did not improve the performance of the previous model (M2) to a statistically significant degree ($z=1.2$, $P=.19$); however, there was a significant performance improvement between M2 and M3 in the training cohort outcomes ($z=0.02$, $P=.02$; see [Multimedia Appendix 3](#)). Other model performance outcomes of M1, M2, and M3 for the holdout cohort were similar to those of the training cohort ([Multimedia Appendix 3](#)), which cross-validates the algorithm. The receiver operating characteristic curves and precision recall curves of the 3 model iterations are charted for comparison in [Figure 1](#). [Figure 2](#) compares the predicted outcomes of M1, M2, and M3 against the actual 2018 mortality rate for those patients in the top decile of predicted mortality likelihood. As features were added with each model iteration, classification of the highest risk members improved. The final model (M3) was superior to both M1 and M2, predicting that those in the top 1% of highest risk would have a mortality rate of 47.4% in 2018 (versus an actual mortality rate of 44.2%).

[Table 3](#) reports the top 20 features and their rank among the 907 total inputs of M3. To aid interpretation, features are categorized by demographics, diagnoses, medical utilization, pharmacy utilization, laboratory utilization, and SDOH. The absolute (unnormalized) impact values of the top 20 features

are shown in Figure 3. Patient demographics (age and gender) were 2 of the inputs comprising M1, and these were also the most important features contributing to the M3 mortality model. Notably, 3 of the top 20 model features quantify patient information from the total claims data set (total claims, average cost of claim, total diagnoses), and 1 feature was strictly temporal (time since last outpatient visit). Among the top features in M3, 4 inputs captured patient diagnoses, with chronic respiratory disease and kidney disease having the greatest ranked importance (#3 and #8, respectively). Aside from age and gender, kidney disease occurrence was the only other input from M1 to rank in the top 20 features of M3. Additionally, 4 of the 265 medical utilization features were also among the top 20, with total patient claims ranking as the most important in the category (#4) followed by the patient's average cost of claim

(#11). Of the 198 pharmacy utilization inputs, 7 ranked in the top 20 features of M3; 3 of these were among the top 10 most important features in the final ML model. These were antihyperlipidemics (#5), furosemide (#7), and anti-inflammatory analgesics (#9). Although there were 201 laboratory utilization inputs, only 1 was among the top 20 most important features in M3 (lipid panel test, #6). The laboratory features were extracted from claims data and only measure utilization; actual results of patient laboratory tests were not a part of the data used to develop the ML model. Finally, 2 of the 8 patient SDI score features ranked among the top 20 features of M3. The important SDOH features predicting mortality in M3 were food access score (#10) and local economy score (#12) based on the plan member's census tract.

Table 1. Sample member characteristics.

Characteristic	Total sample (n=318,774)	Alive (n=306,227, 96.1%)	Deceased (n=12,547, 3.9%)
Gender, n (%)			
Female	181,158 (56.8)	174,640 (57.0)	6518 (51.9)
Male	136,970 (43.0)	130,941 (42.8)	6029 (48.1)
Missing/not available	646 (0.2)	646 (0.2)	0 (0)
Age (years), mean (SD)	70.7 (11.5)	70.4 (11.5)	77.2 (9.7)
Medical diagnoses, n (%)			
Chronic respiratory disease	56,734 (10.4)	52,183 (10.2)	4551 (14.0)
Heart failure	54,702 (10.1)	50,254 (9.8)	4448 (13.7)
Cancer	44,145 (8.1)	40,985 (8.0)	3160 (9.7)
Stroke	21,338 (3.9)	19,327 (3.8)	2011 (6.2)
Dementia or Alzheimer disease	15,626 (2.9)	13,018 (2.5)	2608 (8.0)
Hypertension	204,405 (37.6)	195,035 (38.2)	9370 (28.8)
Diabetes	146,394 (26.9)	139,999 (27.4)	6395 (19.7)
Medical service utilization, mean (SD)			
Total care visits per year ^a	20.8 (39.5)	20.2 (38.2)	36.7 (60.9)
Emergency room visits per year	0.4 (1.1)	0.4 (1.1)	0.9 (1.7)
Pharmacy utilization, mean (SD)			
Total unique medications prescribed	9.04 (7.4)	8.9 (7.3)	11.7 (8.3)
Number of prescribed medications per day	8.11 (12.0)	8.0 (12.1)	9.8 (9.9)
Laboratory utilization, mean (SD)			
Total unique lab-related CPT ^b codes	8.7 (8.4)	8.6 (8.2)	11.7 (11.0)
Social determinants index (SDI)^c, mean (SD)			
Weighted SDI score ^d	58.41 (8.65)	58.43 (8.67)	58.09 (8.08)
Unweighted SDI score ^d	56.94 (10.12)	56.98 (10.13)	55.91 (9.63)

^aIncludes all inpatient and outpatient visits.

^bCPT: Current Procedural Terminology.

^cHigher is better.

^d100 points maximum.

Table 2. Model summary and performance comparison (holdout cohort).

Measure	Model 1 (M1; baseline)	Model 2 (M2)	Model 3 (M3; final)
Total model features, n	7	899	907
Model input summary	Demographics ^a , High-risk utilization indicators ^{b,c}	Demographics ^a , High-risk utilization indicators ^{b,c} ; Medical, lab, and pharmacy utilization ^c	Demographics ^a , High-risk utilization indicators ^{b,c} ; Medical, lab, and pharmacy utilization ^c ; SDI ^d scores ^a
Model performance (holdout cohort)			
AUC ^e (95% CI)	0.736 (0.728-0.744)	0.834 (0.828-0.840)	0.839 (0.833-0.845)
True positive rate ^f	0.105	0.320	0.2993
PPV ^{f,g}	0.212	0.264	0.2991
False positive rate ^f	0.016	0.037	0.029
True negative rate ^f	0.984	0.963	0.97126
NPV ^{f,h}	0.964	0.972	0.97129
False negative rate ^f	0.890	0.679	0.701
AP ⁱ	0.122	0.233	0.243
Performance comparison (holdout cohort)			
Null hypothesis	$AUC_{M1} = 0.5$	$AUC_{M2} - AUC_{M1} = 0.0$	$AUC_{M3} - AUC_{M2} = 0.0$
z statistic	56.4	19.1	1.2
P value	<.001	<.001	.19

^aSource: internal administrative member records.

^bSource: electronic health record (EHR) data.

^cSource: claims data.

^dSDI: social determinants index.

^eAUC: area under the curve.

^fValues based on a defined threshold of 0.16.

^gPPV: positive predictive value.

^hNPV: negative predictive value.

ⁱAP: average precision.

Figure 1. Comparison of Model 1 (M1), Model 2 (M2), and Model 3 (M3) using (A) receiver operating characteristic curves and (B) precision recall curves. AP: average precision; AUC: area under the receiver operating characteristic curve.

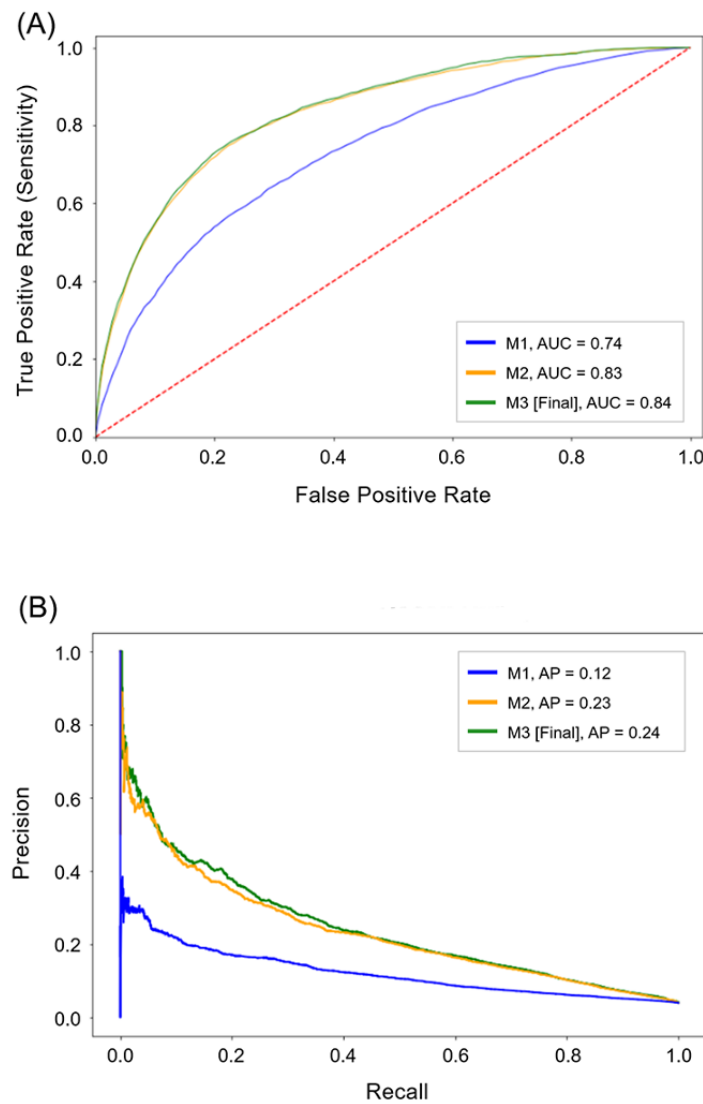


Figure 2. Model mortality outcomes for patients in the top decile of the highest predicted risk. M1: Model 1; M2: Model 2; M3: Model 3.

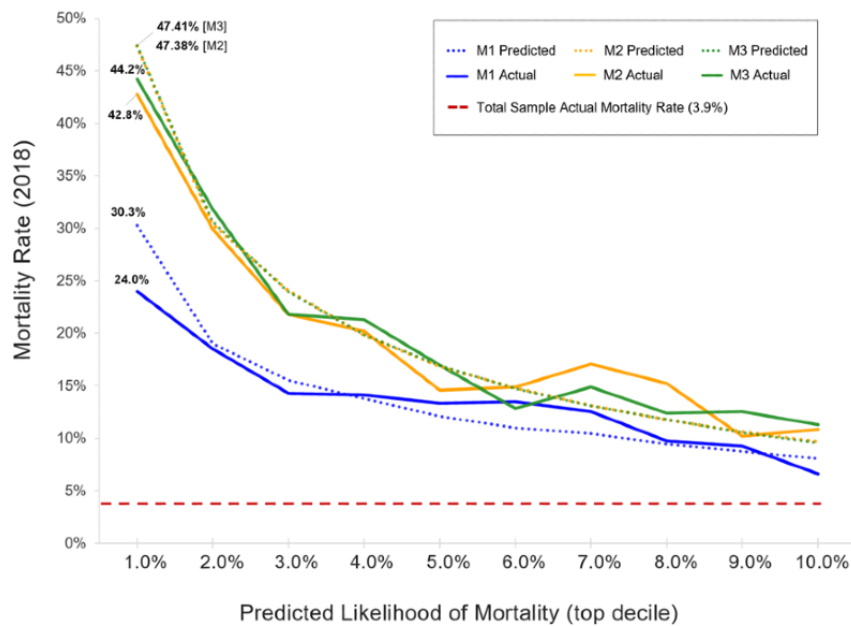


Table 3. Ranked importance of top features in the final model (M3; 907 total inputs).

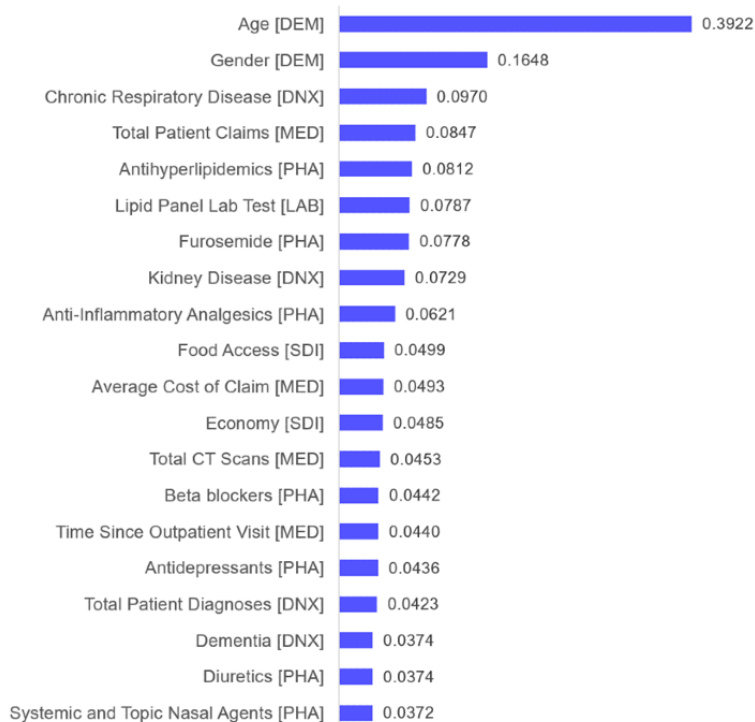
Feature category and M3 features	M3 ranked importance ^a
Demographics (2 inputs)	
Age ^b	1
Gender ^b	2
Diagnoses (233 inputs)	
Chronic respiratory disease	3
Kidney disease ^b	8
Total patient diagnoses	17
Dementia	18
Medical utilization (265 inputs)	
Total patient claims	4
Average cost of claim	11
Total CT ^c scans	13
Time since last outpatient visit	15
Pharmacy utilization (198 inputs)	
Antihyperlipidemics	5
Furosemide	7
Anti-inflammatory analgesics	9
Beta blockers	14
Antidepressants	16
Diuretics	19
Laboratory utilization (201 inputs)	
Systemic and topical nasal agents	20
Lipid panel lab test	6
Social determinants index (SDI) score (8 inputs)	
Food access	10
Economy	12

^aRanked importance based on positive Shapley Additive Explanations value of features.

^bM1 feature.

^cCT: computed tomography.

Figure 3. Absolute feature importance in Model 3 (M3). CT: computed tomography; DEM: demographics; DNX: diagnoses; LAB: laboratory utilization; MED: medical utilization; PHA: pharmacy utilization; SDI: social determinants index.



Discussion

Principal Findings

In the past, provider groups and physicians have relied on manual checking of patient records to prescribe palliative care for patients. Today, palliative care teams are increasingly using enhanced decision tools, such as ML approaches, for expedient care delivery. Our palliative care ML model aims to provide a more objective, automated way to identify patients in Medicare Advantage who could most benefit from palliative services, ensuring appropriate clinical resource allocation to the patients with the highest need. The health plan's goal is to optimize the patient's quality of life outcomes and incorporate all aspects of palliative care, including care coordination, polypharmacy, symptom management, advanced care plans, as well as spiritual and psychosocial assessments. In this sense, identifying patients who can benefit from a palliative care intervention takes a whole-person health approach to chronic health management and end of life care; the focus is not solely on a transition to hospice. In practice, the model could be deployed within case management, home health, or direct-to-provider programs.

Earlier ML studies of community-dwelling older Medicare beneficiaries have attempted to refine the predictive capabilities of various ML model configurations. However, few have reported outcomes of their specific model feature inputs [25-29]. Understanding important features contributing to mortality prediction algorithms can highlight differences in outcomes between models based on the population studied, ML model approach, and type of data analyzed. Increased transparency in reporting model feature outcomes may also help inform the criterion validity of existing clinical assessment tools used to evaluate patients for palliative care needs. Furthermore, features

capturing the SDOH have also been largely neglected from ML models in previous literature [6,25-29,31-33,36,41]. Our feature impact outcomes show that SDOH (ie, food access and local economy) not only are relevant to the prediction of end of life in the community-dwelling Medicare Advantage population but also may be more influential on the outcome than some archetypal high-risk diagnostic and service utilization indicators of palliative care need that are perhaps more commonly observed in hospital settings (eg, ventilator use, sepsis).

The performance of our baseline gradient-boosted machine model predicting 1-year mortality in Medicare Advantage plan members (aged ≥ 65 years) improved with the incorporation of patient service utilization, diagnoses, and SDOH features. Having access to and adding the full medical, laboratory, and pharmacy claims data and SDI measures enhanced our ML approach. The performance of our model is comparable to that of previous ML studies of older community-dwelling Medicare beneficiaries using claims data (see [Multimedia Appendix 4](#)). Some of these models have achieved greater accuracy than that in this study, particularly those models using deep learning configurations. For example, the long short-term memory and deep neural network models developed by Guo et al [25] outperformed their random forest model for predicting mortality in outpatients. Although these types of ML models may achieve greater accuracy, the enhanced model complexity and types of data analyzed by deep learning configurations may not be available or necessary in some cases. Patient medical claims are a common and plentiful source of data that can be used to train binary classification ML algorithms for predicting mortality and other health outcomes. In contrast to inputs already defined within discrete data sets, model inputs generated from raw text might also produce more ambiguous feature definitions that could create challenges for feature impact reporting.

Classification models using routine, standard data (ie, claims, administrative records) may be a more attractive option for health plans and other organizations that already collect such data with predefined discrete variables to fulfill their business purposes.

Limitations

Age and gender were the most influential features in our final model. Although these demographic features had substantial impact on the mortality risk outcome, it is unsurprising that age is the most important model feature, as the probability of death increases with age in older individuals. There is also evidence that, for various reasons, men may be likelier to die earlier than women [42]. The importance of age as a predictive variable is documented in the feature reporting of studies on ML mortality models for hospitalized patients [43]. For community-dwelling Medicare Advantage members over 65 years of age, omitting the age or gender inputs may influence the prediction of mortality risk in cases for which the outcome could be better explained by these demographic variables. Race and ethnicity were purposefully excluded from the model. Race and ethnicity are related to certain disease outcomes, but the literature suggests that social determinants may mediate or modify observed racial or ethnic health differences [44]. When predicting mortality, we believe the composite SDI scores provide more information on the regional variation in individual levels of SDOH and potentially less measurement bias compared with patient race or ethnicity [33].

Our model was developed using only data from a nationwide population sample of community-dwelling Medicare Advantage

plan members aged 65 years or older, which could constrain the generalizability of study results to other kinds of patient groups and health settings. Although our model was trained based just on the Medicare Advantage population, bidirectional data sharing between US commercial and other government products would allow for other types of health care consumers to benefit from ML tools for early identification of patients for palliative care. Additionally, our ML model was built to be generic and disease-agnostic. The mortality outcome for the year 2018 encompassed all causes of death, and the feature generation period was also randomized with the span of 1 year. Although the model's applicability to different patient populations and care settings is still unknown, the generic model can be applied to the plan's Medicare Advantage members across different years.

Conclusion

ML offers greater precision and sensitivity in predicting patient end of life and potential need for palliative services among community-dwelling older Medicare beneficiaries. In response to a lack of feature reporting in relevant previous research, this study explored the development of a binary classification ML algorithm predicting 1-year mortality among a sample of Medicare Advantage plan members and investigated the mortality model's features of top importance. We found the most important features included demographics, diagnoses, pharmacy utilization, mean costs, and certain SDOH. The final ML model predicts mortality among Medicare Advantage plan members with a high degree of accuracy and precision using a variety of routinely collected data and can support earlier patient identification for palliative care.

Acknowledgments

The authors would like to acknowledge and thank Joshua Barrett and Dr Mayank Shah for their important contributions to the development of this manuscript.

Conflicts of Interest

AB, CD, AEM, RM, and AT are employees of the organization that requested and funded the study (Cigna/Evernorth). BM is a contracted employee of the same organization. The authors have no further interests to declare.

Multimedia Appendix 1

Health plan process for identifying palliative care patients using machine learning.

[\[DOCX File, 94 KB - ai_v2i1e42253_app1.docx\]](#)

Multimedia Appendix 2

Social determinants index (SDI) select measures summary.

[\[DOCX File, 19 KB - ai_v2i1e42253_app2.docx\]](#)

Multimedia Appendix 3

Model summary and performance comparison (Training Cohort).

[\[DOCX File, 19 KB - ai_v2i1e42253_app3.docx\]](#)

Multimedia Appendix 4

Machine learning (ML) models predicting patient mortality for earlier identification for palliative care.

[\[DOCX File, 23 KB - ai_v2i1e42253_app4.docx\]](#)

References

1. March 2021 Report to the Congress: Medicare Payment Policy. Medicare Payment Advisory Commission. 2021. URL: <https://www.medpac.gov/document/march-2021-report-to-the-congress-medicare-payment-policy/> [accessed 2023-01-06]
2. May P, Tysinger B, Morrison RS, Jacobson M. Advancing the economics of palliative care: The value to individuals and families, organizations, and society. USC Schaeffer Center for Health Policy & Economics. 2021 Aug 5. URL: <https://healthpolicy.usc.edu/research/advancing-the-economics-of-palliative-care-the-value-to-individuals-and-families-organizations-and-society/> [accessed 2023-01-06]
3. Bevins J, Bhulani N, Goksu SY, Sanford NN, Gao A, Ahn C, et al. Early palliative care is associated with reduced emergency department utilization in pancreatic cancer. *Am J Clin Oncol* 2021 May 01;44(5):181-186. [doi: [10.1097/COC.0000000000000802](https://doi.org/10.1097/COC.0000000000000802)] [Medline: [33710133](https://pubmed.ncbi.nlm.nih.gov/33710133/)]
4. Cunningham C, Ollendorf D, Travers K. The effectiveness and value of palliative care in the outpatient setting. *JAMA Intern Med* 2017 Feb 01;177(2):264-265. [doi: [10.1001/jamainternmed.2016.8177](https://doi.org/10.1001/jamainternmed.2016.8177)] [Medline: [28055045](https://pubmed.ncbi.nlm.nih.gov/28055045/)]
5. De Jonge KE, Jamshed N, Gilden D, Kubisiak J, Bruce SR, Taler G. Effects of home-based primary care on Medicare costs in high-risk elders. *J Am Geriatr Soc* 2014 Oct 18;62(10):1825-1831. [doi: [10.1111/jgs.12974](https://doi.org/10.1111/jgs.12974)] [Medline: [25039690](https://pubmed.ncbi.nlm.nih.gov/25039690/)]
6. Zhang B, Wright AA, Huskamp HA, Nilsson ME, Maciejewski ML, Earle CC, et al. Health care costs in the last week of life: Associations with end-of-life conversations. *Arch Intern Med* 2009 Mar 09;169(5):480-488. [doi: [10.1001/archinternmed.2008.587](https://doi.org/10.1001/archinternmed.2008.587)] [Medline: [19273778](https://pubmed.ncbi.nlm.nih.gov/19273778/)]
7. Vallabhajosyula S, Prasad A, Dunlay SM, Murphree DH, Ingram C, Mueller PS, et al. Utilization of palliative care for cardiogenic shock complicating acute myocardial infarction: A 15 - year national perspective on trends, disparities, predictors, and outcomes. *JAHA* 2019 Aug 06;8(15):e011954. [doi: [10.1161/jaha.119.011954](https://doi.org/10.1161/jaha.119.011954)]
8. Seow H, O'Leary E, Perez R, Tanuseputro P. Access to palliative care by disease trajectory: A population-based cohort of Ontario decedents. *BMJ Open* 2018 Apr 05;8(4):e021147 [FREE Full text] [doi: [10.1136/bmjopen-2017-021147](https://doi.org/10.1136/bmjopen-2017-021147)] [Medline: [29626051](https://pubmed.ncbi.nlm.nih.gov/29626051/)]
9. Etkind SN, Bone AE, Gomes B, Lovell N, Evans CJ, Higginson IJ, et al. How many people will need palliative care in 2040? Past trends, future projections and implications for services. *BMC Med* 2017 May 18;15(1):102 [FREE Full text] [doi: [10.1186/s12916-017-0860-2](https://doi.org/10.1186/s12916-017-0860-2)] [Medline: [28514961](https://pubmed.ncbi.nlm.nih.gov/28514961/)]
10. Yen Y, Hu H, Lai Y, Chou Y, Chen C, Ho C. Comparison of intuitive assessment and palliative care screening tool in the early identification of patients needing palliative care. *Sci Rep* 2022 Mar 23;12(1):4955 [FREE Full text] [doi: [10.1038/s41598-022-08886-7](https://doi.org/10.1038/s41598-022-08886-7)] [Medline: [35322098](https://pubmed.ncbi.nlm.nih.gov/35322098/)]
11. Downar J, Wegier P, Tanuseputro P. Early identification of people who would benefit from a palliative approach—Moving from surprise to routine. *JAMA Netw Open* 2019 Sep 04;2(9):e1911146 [FREE Full text] [doi: [10.1001/jamanetworkopen.2019.11146](https://doi.org/10.1001/jamanetworkopen.2019.11146)] [Medline: [31517959](https://pubmed.ncbi.nlm.nih.gov/31517959/)]
12. White N, Reid F, Harris A, Harries P, Stone P. A systematic review of predictions of survival in palliative care: How accurate are clinicians and who are the experts? *PLoS One* 2016 Aug 25;11(8):e0161407 [FREE Full text] [doi: [10.1371/journal.pone.0161407](https://doi.org/10.1371/journal.pone.0161407)] [Medline: [27560380](https://pubmed.ncbi.nlm.nih.gov/27560380/)]
13. Fischer SM, Gozansky WS, Sauaia A, Min S, Kutner JS, Kramer A. A practical tool to identify patients who may benefit from a palliative approach: The CARING criteria. *J Pain Symptom Manage* 2006 Apr;31(4):285-292 [FREE Full text] [doi: [10.1016/j.jpainsymman.2005.08.012](https://doi.org/10.1016/j.jpainsymman.2005.08.012)] [Medline: [16632076](https://pubmed.ncbi.nlm.nih.gov/16632076/)]
14. Pilsworth S, Wat D, Sibley S, Crane J. A service evaluation of the accuracy of the Gold Standard Framework Proactive Indicator Guidance (GSF PIG) in predicting 12 month mortality in patients with a diagnosis of chronic obstructive pulmonary disease (COPD). *Eur Respir J* 2018 Nov 19;52(suppl 62):PA3859. [doi: [10.1183/13993003.congress-2018.PA3859](https://doi.org/10.1183/13993003.congress-2018.PA3859)]
15. Gómez-Batiste X, Turrillas P, Tebé C, Calsina-Berna A, Amblàs-Novellas J. NECPAL tool prognostication in advanced chronic illness: A rapid review and expert consensus. *BMJ Support Palliat Care* 2022 May 02;12(e1):e10-e20. [doi: [10.1136/bmjspcare-2019-002126](https://doi.org/10.1136/bmjspcare-2019-002126)] [Medline: [32241958](https://pubmed.ncbi.nlm.nih.gov/32241958/)]
16. Wharton T, Manu E, Vitale CA. Enhancing provider knowledge and patient screening for palliative care needs in chronic multimorbid patients receiving home-based primary care. *Am J Hosp Palliat Care* 2015 Feb 25;32(1):78-83. [doi: [10.1177/1049909113514475](https://doi.org/10.1177/1049909113514475)] [Medline: [24280188](https://pubmed.ncbi.nlm.nih.gov/24280188/)]
17. Thoonsen B, Engels Y, van Rijswijk E, Verhagen S, van Weel C, Groot M, et al. Early identification of palliative care patients in general practice: Development of RADboud indicators for Palliative Care Needs (RADPAC). *Br J Gen Pract* 2012 Sep 01;62(602):e625-e631. [doi: [10.3399/bjgp12x654597](https://doi.org/10.3399/bjgp12x654597)]
18. Hight G, Crawford D, Murray SA, Boyd K. Development and evaluation of the Supportive and Palliative Care Indicators Tool (SPICe): A mixed-methods study. *BMJ Support Palliat Care* 2014 Sep 25;4(3):285-290. [doi: [10.1136/bmjspcare-2013-000488](https://doi.org/10.1136/bmjspcare-2013-000488)] [Medline: [24644193](https://pubmed.ncbi.nlm.nih.gov/24644193/)]
19. ElMokhallalati Y, Bradley SH, Chapman E, Ziegler L, Murtagh FE, Johnson MJ, et al. Identification of patients with potential palliative care needs: A systematic review of screening tools in primary care. *Palliat Med* 2020 Sep 07;34(8):989-1005 [FREE Full text] [doi: [10.1177/0269216320929552](https://doi.org/10.1177/0269216320929552)] [Medline: [32507025](https://pubmed.ncbi.nlm.nih.gov/32507025/)]

20. Walsh RI, Mitchell G, Francis L, van Driel ML. What diagnostic tools exist for the early identification of palliative care patients in general practice? A systematic review. *J Palliat Care* 2015 Nov 13;31(2):118-123. [doi: [10.1177/082585971503100208](https://doi.org/10.1177/082585971503100208)] [Medline: [26201214](https://pubmed.ncbi.nlm.nih.gov/26201214/)]
21. Davies JM, Gao W, Sleeman KE, Lindsey K, Murtagh FE, Teno JM, et al. Using routine data to improve palliative and end of life care. *BMJ Support Palliat Care* 2016 Sep 28;6(3):257-262. [doi: [10.1136/bmjspcare-2015-000994](https://doi.org/10.1136/bmjspcare-2015-000994)] [Medline: [26928173](https://pubmed.ncbi.nlm.nih.gov/26928173/)]
22. Storick V, O'Herlihy A, Abdelhafeez S, Ahmed R, May P. Improving palliative and end-of-life care with machine learning and routine data: A rapid review. *HRB Open Res* 2019 Aug 12;2:13 [FREE Full text] [doi: [10.12688/hrbopenres.12923.2](https://doi.org/10.12688/hrbopenres.12923.2)] [Medline: [32002512](https://pubmed.ncbi.nlm.nih.gov/32002512/)]
23. Austin SR, Wong Y, Uzzo RG, Beck JR, Egleston BL. Why summary comorbidity measures such as the Charlson Comorbidity Index and Elixhauser Score work. *Med Care* 2015 Sep;53(9):e65-e72 [FREE Full text] [doi: [10.1097/MLR.0b013e318297429c](https://doi.org/10.1097/MLR.0b013e318297429c)] [Medline: [23703645](https://pubmed.ncbi.nlm.nih.gov/23703645/)]
24. Gagne JJ, Glynn RJ, Avorn J, Levin R, Schneeweiss S. A combined comorbidity score predicted mortality in elderly patients better than existing scores. *J Clin Epidemiol* 2011 Jul;64(7):749-759. [doi: [10.1016/j.jclinepi.2010.10.004](https://doi.org/10.1016/j.jclinepi.2010.10.004)] [Medline: [21208778](https://pubmed.ncbi.nlm.nih.gov/21208778/)]
25. Guo A, Foraker R, White P, Chivers C, Courtright K, Moore N. Using electronic health records and claims data to identify high-risk patients likely to benefit from palliative care. *Am J Manag Care* 2021 Jan 01;27(1):e7-e15 [FREE Full text] [doi: [10.37765/ajmc.2021.88578](https://doi.org/10.37765/ajmc.2021.88578)] [Medline: [33471463](https://pubmed.ncbi.nlm.nih.gov/33471463/)]
26. Berg GD, Gurley VF. Development and validation of 15-month mortality prediction models: A retrospective observational comparison of machine-learning techniques in a national sample of Medicare recipients. *BMJ Open* 2019 Jul 16;9(7):e022935 [FREE Full text] [doi: [10.1136/bmjopen-2018-022935](https://doi.org/10.1136/bmjopen-2018-022935)] [Medline: [31315852](https://pubmed.ncbi.nlm.nih.gov/31315852/)]
27. Makar M, Ghassemi M, Cutler DM, Obermeyer Z. Short-term mortality prediction for elderly patients using Medicare claims data. *IJMLC* 2015 Jun;5(3):192-197. [doi: [10.7763/ijmlc.2015.v5.506](https://doi.org/10.7763/ijmlc.2015.v5.506)]
28. Hamlet KS, Hobgood A, Hamar GB, Dobbs AC, Rula EY, Pope JE. Impact of predictive model-directed end-of-life counseling for Medicare beneficiaries. *Am J Manag Care* 2010 May;16(5):379-384 [FREE Full text] [Medline: [20469958](https://pubmed.ncbi.nlm.nih.gov/20469958/)]
29. Cary MP, Zhuang F, Draelos RL, Pan W, Amarasekara S, Douthit BJ, et al. Machine learning algorithms to predict mortality and allocate palliative care for older patients with hip fracture. *J Am Med Dir Assoc* 2021 Feb;22(2):291-296. [doi: [10.1016/j.jamda.2020.09.025](https://doi.org/10.1016/j.jamda.2020.09.025)] [Medline: [33132014](https://pubmed.ncbi.nlm.nih.gov/33132014/)]
30. Feature Impact. DataRobot. 2022 Nov 8. URL: <https://docs.datarobot.com/en/docs/modeling/analyze-models/understand/feature-impact.html> [accessed 2023-01-06]
31. Avati A, Jung K, Harman S, Downing L, Ng A, Shah NH. Improving palliative care with deep learning. *BMC Med Inform Decis Mak* 2018 Dec 12;18(suppl 4):122 [FREE Full text] [doi: [10.1186/s12911-018-0677-8](https://doi.org/10.1186/s12911-018-0677-8)] [Medline: [30537977](https://pubmed.ncbi.nlm.nih.gov/30537977/)]
32. Sahni N, Simon G, Arora R. Development and validation of machine learning models for prediction of 1-year mortality utilizing electronic medical record data available at the end of hospitalization in multicondition patients: A proof-of-concept study. *J Gen Intern Med* 2018 Jun 30;33(6):921-928 [FREE Full text] [doi: [10.1007/s11606-018-4316-y](https://doi.org/10.1007/s11606-018-4316-y)] [Medline: [29383551](https://pubmed.ncbi.nlm.nih.gov/29383551/)]
33. Agarwal R, Domenico HJ, Balla SR, Byrne DW, Whisenant JG, Woods MC, et al. Palliative care exposure relative to predicted risk of six-month mortality in hospitalized adults. *J Pain Symptom Manage* 2022 May;63(5):645-653. [doi: [10.1016/j.jpainsymman.2022.01.013](https://doi.org/10.1016/j.jpainsymman.2022.01.013)] [Medline: [35081441](https://pubmed.ncbi.nlm.nih.gov/35081441/)]
34. Hilal S, Brayne C. Epidemiologic trends, social determinants, and brain health: The role of life course inequalities. *Stroke* 2022 Feb;53(2):437-443. [doi: [10.1161/strokeaha.121.032609](https://doi.org/10.1161/strokeaha.121.032609)]
35. Silva VDL, Cesse EP, de Albuquerque MFPM. Social determinants of death among the elderly: A systematic literature review. *Rev Bras Epidemiol* 2014;17(suppl 2):178-193 [FREE Full text] [doi: [10.1590/1809-4503201400060015](https://doi.org/10.1590/1809-4503201400060015)] [Medline: [25409647](https://pubmed.ncbi.nlm.nih.gov/25409647/)]
36. Shi Y, Wu Z, Zhang S, Xiao H, Zhao Y. Assessing palliative care needs using machine learning approaches. 2021 Presented at: 45th Annual Computers, Software, and Applications Conference (COMPSAC); July 12-16, 2021; Madrid, Spain. [doi: [10.1109/compsac51774.2021.00049](https://doi.org/10.1109/compsac51774.2021.00049)]
37. Luo W, Phung D, Tran T, Gupta S, Rana S, Karmakar C, et al. Guidelines for developing and reporting machine learning predictive models in biomedical research: A multidisciplinary view. *J Med Internet Res* 2016 Dec 16;18(12):e323 [FREE Full text] [doi: [10.2196/jmir.5870](https://doi.org/10.2196/jmir.5870)] [Medline: [27986644](https://pubmed.ncbi.nlm.nih.gov/27986644/)]
38. Checking Medicare Eligibility. Medicare Learning Network. Baltimore, MD: Centers for Medicare & Medicaid Services; 2022 Oct. URL: <https://www.cms.gov/files/document/checking-medicare-eligibility.pdf> [accessed 2023-01-07]
39. American Community Survey Data via FTP. United States Census Bureau. Suitland-Silver Hill, MD: United States Census Bureau; 2022. URL: <https://www.census.gov/programs-surveys/acs/data/data-via-ftp.html> [accessed 2023-01-07]
40. Lundberg SM, Lee SI. A unified approach to interpreting model predictions. In: Guyon I, Von Luxburg U, Bengio S, Wallach H, Fergus R, Vishwanathan S, et al, editors. *Advances in Neural Information Processing Systems 30 (NIPS 2017)*. La Jolla, CA: Neural Information Process Systems; 2017:1-10.
41. Zhang H, Li Y, McConnell W. Predicting potential palliative care beneficiaries for health plans: A generalized machine learning pipeline. *J Biomed Inform* 2021 Nov;123:103922 [FREE Full text] [doi: [10.1016/j.jbi.2021.103922](https://doi.org/10.1016/j.jbi.2021.103922)] [Medline: [34607012](https://pubmed.ncbi.nlm.nih.gov/34607012/)]

42. Regan JC, Partridge L. Gender and longevity: Why do men die earlier than women? Comparative and experimental evidence. *Best Pract Res Clin Endocrinol Metab* 2013 Aug;27(4):467-479. [doi: [10.1016/j.beem.2013.05.016](https://doi.org/10.1016/j.beem.2013.05.016)] [Medline: [24054925](https://pubmed.ncbi.nlm.nih.gov/24054925/)]
43. Park JY, Hsu T, Hu J, Chen C, Hsu W, Lee M, et al. Predicting sepsis mortality in a population-based national database: Machine learning approach. *J Med Internet Res* 2022 Apr 13;24(4):e29982 [FREE Full text] [doi: [10.2196/29982](https://doi.org/10.2196/29982)] [Medline: [35416785](https://pubmed.ncbi.nlm.nih.gov/35416785/)]
44. Lorch SA, Enlow E. The role of social determinants in explaining racial/ethnic disparities in perinatal outcomes. *Pediatr Res* 2016 Jan 14;79(1):141-147. [doi: [10.1038/pr.2015.199](https://doi.org/10.1038/pr.2015.199)] [Medline: [26466077](https://pubmed.ncbi.nlm.nih.gov/26466077/)]

Abbreviations

AUC: area under the receiver operating characteristic curve

CPT: Current Procedural Terminology

EHR: electronic health record

LightGBM: light gradient-boosted tree model

M1: Model 1

M2: Model 2

M3: Model 3

ML: machine learning

SDI: social determinants index

SDOH: social determinants of health

SHAP: Shapley Additive Explanations

Edited by K El Emam, B Malin; submitted 29.08.22; peer-reviewed by L Juwara, J Li; comments to author 26.09.22; revised version received 21.11.22; accepted 20.12.22; published 20.02.23.

Please cite as:

Bowers A, Drake C, Makarkin AE, Monzyk R, Maity B, Telle A

Predicting Patient Mortality for Earlier Palliative Care Identification in Medicare Advantage Plans: Features of a Machine Learning Model

JMIR AI 2023;2:e42253

URL: <https://ai.jmir.org/2023/1/e42253>

doi: [10.2196/42253](https://doi.org/10.2196/42253)

PMID:

©Anne Bowers, Chelsea Drake, Alexi E Makarkin, Robert Monzyk, Biswajit Maity, Andrew Telle. Originally published in JMIR AI (<https://ai.jmir.org>), 20.02.2023. This is an open-access article distributed under the terms of the Creative Commons Attribution License (<https://creativecommons.org/licenses/by/4.0/>), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR AI, is properly cited. The complete bibliographic information, a link to the original publication on <https://www.ai.jmir.org/>, as well as this copyright and license information must be included.

Original Paper

Deep Learning to Detect Pancreatic Cystic Lesions on Abdominal Computed Tomography Scans: Development and Validation Study

Maria Montserrat Duh^{1*}, MD; Neus Torra-Ferrer^{1*}, MD; Meritxell Riera-Marín², MSc; Dídac Cumelles², BSc; Júlia Rodríguez-Comas², PhD; Javier García López², PhD; M^a Teresa Fernández Planas^{1*}, MD

¹Department of Radiology, Consorci Sanitari del Maresme (Hospital de Mataró), Mataró, Spain

²Scientific and Technical Department, Sycal Technologies SL, Barcelona, Spain

*these authors contributed equally

Corresponding Author:

M^a Teresa Fernández Planas, MD

Department of Radiology

Consorci Sanitari del Maresme (Hospital de Mataró)

Carretera de Cirera, 230

Mataró, 08304

Spain

Phone: 34 674152399 ext 754

Email: mfernandezpl@csdm.cat

Abstract

Background: Pancreatic cystic lesions (PCLs) are frequent and underreported incidental findings on computed tomography (CT) scans and can evolve to pancreatic cancer—the most lethal cancer, with less than 5 months of life expectancy.

Objective: The aim of this study was to develop and validate an artificial deep neural network (attention gate U-Net, also named “AGNet”) for automated detection of PCLs. This kind of technology can help radiologists to cope with an increasing demand of cross-sectional imaging tests and increase the number of PCLs incidentally detected, thus increasing the early detection of pancreatic cancer.

Methods: We adapted and evaluated an algorithm based on an attention gate U-Net architecture for automated detection of PCL on CTs. A total of 335 abdominal CTs with PCLs and control cases were manually segmented in 3D by 2 radiologists with over 10 years of experience in consensus with a board-certified radiologist specialized in abdominal radiology. This information was used to train a neural network for segmentation followed by a postprocessing pipeline that filtered the results of the network and applied some physical constraints, such as the expected position of the pancreas, to minimize the number of false positives.

Results: Of 335 studies included in this study, 297 had a PCL, including serous cystadenoma, intraductal pseudopapillary mucinous neoplasia, mucinous cystic neoplasm, and pseudocysts. The Shannon Index of the chosen data set was 0.991 with an evenness of 0.902. The mean sensitivity obtained in the detection of these lesions was 93.1% (SD 0.1%), and the specificity was 81.8% (SD 0.1%).

Conclusions: This study shows a good performance of an automated artificial deep neural network in the detection of PCL on both noncontrast- and contrast-enhanced abdominal CT scans.

(JMIR AI 2023;2:e40702) doi:[10.2196/40702](https://doi.org/10.2196/40702)

KEYWORDS

deep learning; pancreatic cystic lesion; neural networks; precursor lesions; pancreatic cancer; computed tomography; magnetic resonance; cancer; radiologist; technology

Introduction

Pancreatic cancer is one of the most frequent and aggressive cancers in the digestive tract, being the fourth leading cause of death by cancer in Europe [1,2]. Due to its lack of specific symptoms and signs, most patients are detected in an advanced

stage. The current average 5-year survival rate is 9%, and it depends critically on when the cancer is detected. Indeed, this 5-year survival rate varies by more than 30% when the cancer is detected in a phase where it can still be surgically removed and when the cancer has already spread to other tissues in the body [3].

This type of cancer can originate from precursor cystic lesions [4]. Pancreatic cystic lesions (PCL) are increasingly common incidental findings on abdominal imaging tests. Studies have shown that up to 70% of PCLs are diagnosed incidentally on computed tomography (CT) scans due to unrelated symptoms, making CT scans the first accessible source of information. These previously undetected cystic lesions are found on 3% of abdominal CT examinations [5,6] and 13%-21% of abdominal magnetic resonance imaging studies [7,8]. However, autopsy studies have evidenced a much higher prevalence, revealing that up to 50% of the older population may present at least one pancreatic cyst [6].

PCLs have a wide diversity, and their differential diagnosis includes nonneoplastic cysts (pseudocysts) and neoplastic ones. Neoplastic lesions encompass benign lesions, such as serous cystadenomas (SCA), to mucinous lesions, such as mucinous cystic neoplasms (MCN), and intraductal papillary mucinous neoplasm (IPMN), which may progress to PC. Therefore, identifying precancerous mucin-producing cysts offers a unique opportunity for early detection and prevention of PC. Once a PCL is found, patients are recommended to follow up a lifelong surveillance program with imaging modalities (magnetic resonance imaging or CT) to identify early-stage cancer or high-grade dysplasia [9,10]. Consequently, correct management of PCL may prevent progression to pancreatic cancer, while reducing the need for lifelong screening and related costs.

In this complex scenario, automated detection of pancreatic precursor lesions could increase the detection of this underreported entity and help with a proper surveillance of these patients. A limited number of publications regarding this topic have been released in recent years, most of them in an experimental offline setting and applying different methodologies [11]. Additionally, although existing methods of automated analysis have shown to be accurate for images of individual organs, they still struggle to deal with the variability of structures, shape, and location of abdominal organs [12]. Artificial intelligence (AI)-based algorithms have shown promising results in the detection of preneoplastic lesions in the pancreas [13,14], but they are still far from implementation in the clinical practice.

The aim of this study was to develop and test an artificial deep neural network (AGNet) [15] for automated detection of PCLs. This kind of technology can help radiologists to cope with an increasing demand of cross-sectional imaging tests and increase the number of PCLs incidentally detected, thus increasing the early detection of pancreatic cancer.

Methods

Ethical Considerations

Our research adhered to the ethical principles outlined in the 1975 Declaration of Helsinki. The data used in this study were

retrospective and anonymized. The study was approved by the hospital Institutional Ethical Review Board under code 90/20 as an observational retrospective single-center study, and the requirement for informed consent was waived.

Study Population

A total of 297 abdominal, thoracoabdominal, or pelvic CT scans acquired at Hospital de Mataró between 2010 and 2021 and diagnosed with a PCL as well as 38 CT scans as controls were selected for the study. All CT scan images were subjectively checked for quality and absence of relevant respiratory artifacts, which could cause misdiagnosis in the abdominal region. The exclusion criteria were underaged patients, artifacts or bad quality in the CT scan image, and patients having undergone surgery in the past to treat the PCL and having a prosthesis in the pancreas that affects the image. Importantly, patients diagnosed with pancreatic adenocarcinoma or any kind of tumor in the pancreas were also excluded from the study.

Of note, a CT image is considered “bad quality” if there is movement or blurriness in it (mostly in the abdominal area, where the pancreas is located). Studies that included these types of images were excluded from the training and testing set because they would impact the learning process of the network or the testing in a negative way, which could then lead to false negatives or false positives.

The final study population consisted of 136 patients: 73 male (178 studies; mean age 67.75, SD 10.74 years) and 63 female (157 studies; mean age 73.52, SD 10.67 years). A mean of 2 (SD 1.4) CT studies and a median of 2.4 studies were available per patient.

Patients' Characteristics

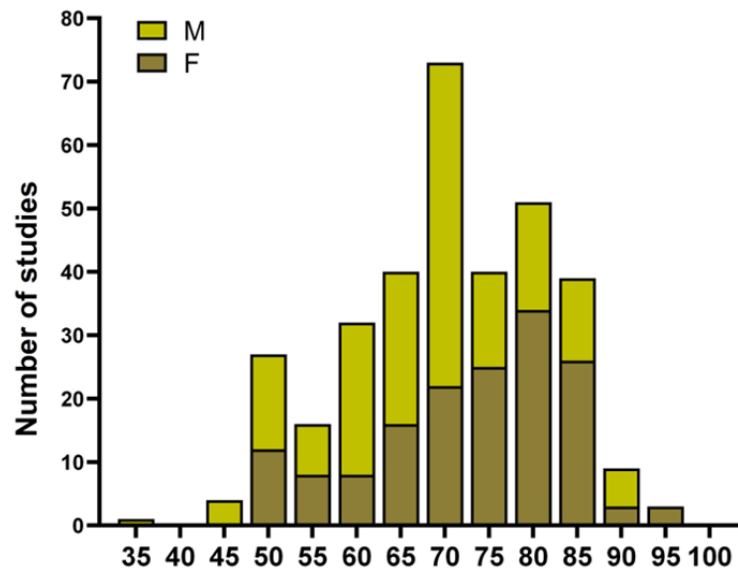
From the whole cohort of 136 patients, 9 (6.5%) of them had a confirmed diagnosis through endoscopic ultrasound-guided fine needle aspiration or surgical resection of the lesion. In the other 16 patients, no material or insufficient yield was extracted to evaluate the specimen. The rest of the patients were diagnosed by a minimum of 2 experienced radiologists, taking into consideration the complete clinical record and the evolution of the patient.

Patients with the following PCLs were included in the study: IPMN, MCN, SCA, and pseudocysts. A total of 14 (4.2%) of the lesions were not classified in the above classification due to unspecified imaging characteristics and were categorized as cyst (Table 1). The number of studies (CT scans) with PCLs distributed by age and sex is shown in Figure 1.

Data sets were further divided between the training set (a subset to train the model) and the testing set (a subset to test the trained model). The final training data set comprised 93 patients, representing a total of 241 CT scans, and the final testing data set comprised 43 patients, representing a total of 94 CT scans. PCLs were distributed proportionally in both data sets.

Table 1. Diagnostic distribution per the study.

Diagnosis	Values, n (%)
Serous cystadenoma	42 (12.5)
Intraductal papillary mucinous neoplasm	154 (46)
Mucinous cystic neoplasms	5 (1.5)
Pseudocyst	82 (24.5)
Cyst	14 (4.2)
No cyst	38 (11.3)

Figure 1. Number of studies (CT scans) with pancreatic cystic lesion distributed by age and sex (x-axis).

CT Protocols

CT examinations were performed with a GE BrightSpeed 16 slice CT scanner (GE Healthcare). Slice thickness was between 1.25 mm and 5 mm. Mean tube current was 440 mA, and the mean peak kilovoltage was 340 (SD 40) kVp. Contrast agent was administered with injection rates ranging from 2.5 to 3 mL/s, using Omnipaque or iomeron (both 300 mg iodine per mL).

The protocols included in this research had the following characteristics:

- From lung bases to pubic symphysis, 2 helices are made at 30 and 65 seconds after the injection of 100 mL of the solution (30 mL of iodine), preceded and followed by 20 mL of physiological solution.
- Two helices are made from the base of the neck to the lower edge of the liver and from the pulmonary bases to the pubic symphysis after the injection of the exposure value contrast. In this case, 120 mL of solution is injected.
- From lung bases to pubic symphysis, 1 helix is made at 65 seconds after the injection of 100 mL of the solution (30 mL of iodine), preceded and followed by 20 mL of physiological solution.

Image Analysis

CT scan images were exported anonymously in Digital Imaging and Communication on Medicine format from the picture

archiving and communication system of the hospital. Digital Imaging and Communication on Medicine files were converted to Neuroimaging Informatics Technology Initiative files (using dicom2nii software; version from August 4, 2014; University of South Carolina). Two radiologists (NTF and MMD) with 11 and 20 years of experience manually drew, slice by slice, the region of interest, delimiting the pancreatic cysts found in the image using the open-source software 3D Slicer (version 4.11) [16]. Each radiologist segmented all cases used in the study and checked the segmentation performed by the other radiologist. Any discrepancies between the authors were resolved through discussion with the presence of a third reviewer (MTFP), until consensus was reached.

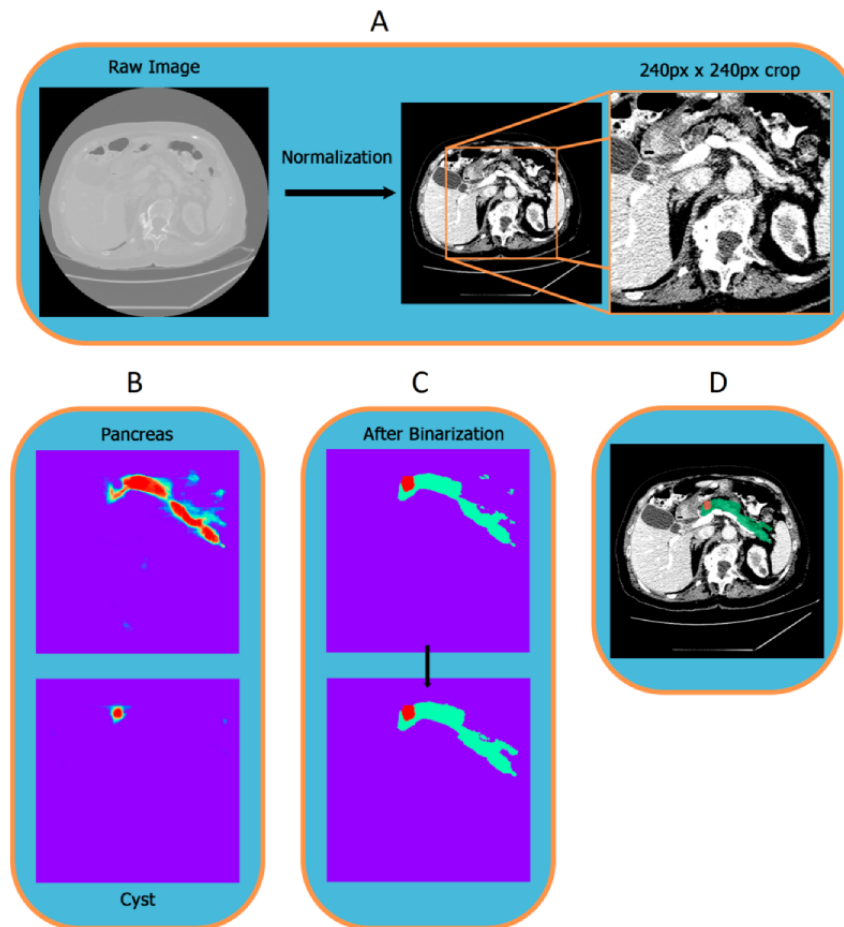
The preprocessing steps included the application of filters and registration to improve and harmonize image quality across CT scans.

First, a soft-tissue normalization [17] was applied. After studying the pixel distribution of 100 CTs of the data set, it was observed and confirmed by the state of the art that the Hounsfield unit (HU) of the pancreas is centered around 50, and most of the cystic lesions were close to this value as well. Hence, to eliminate the irrelevant parts of the abdomen and highlight the main features for the study, the soft-tissue normalization was centered in 50 HU, and a windowing length of ± 100 around 50 HU was applied.

Afterwards, a central cropping of the CTs was performed, only keeping the center of the abdomen, where the pancreas is supposed to be. The cropping was not too harsh to avoid the

possibility of eliminating the pancreas from the CT image being used for the following semantic segmentation study. The image analysis pipeline is depicted in [Figure 2](#).

Figure 2. Diagram of the steps implemented in the pipeline. (A) Preprocessing. (B) Logits. (C) Postprocessing. (D) Output.



Model Training

The neural network used for this study was the AGNet [15]. The main structure was a basic UNet [18] with skip connections and additive attention gates (AGs). The input image was downsampled, using max-pooling, by factor 2 at each scale in the encoding part and trilinearly upsampled by the same factor in the decoding part. In each stage of the encoding-decoding architecture, a skip connection from the corresponding encoding stage to the corresponding decoding stage was added. This skip connection enters to the AG together with the output of the previous decoding stage. Thanks to this skip connections using coarser information, we are able to model the location and the relationship between tissues at a global scale. The architecture of the AG is shown in [Figure 3](#).

The output of these AGs was the element-wise multiplication of the attention coefficients (α) and the input feature maps came from the previous stage of the decoding part (x ; [Figure 4](#)). Attention coefficients were used to identify salient regions and preserve only activations that are relevant. There is one

attention coefficient computed for each pixel vector \boxed{x} , where F^l corresponds to the feature maps in layer l . In the case of this study, there are multidimensional attention coefficients, each dimension corresponding to one class. The other input of the AG was a gating vector \boxed{g} , which contained contextual information to determine focus regions. The AGs used were additive since addition between the gating signal and the feature maps were used to obtain the attention coefficients.

The network was trained for 700 epochs and had a batch size of 4. The training was performed with over 430 3D CT studies. The algorithm of optimization used was Adam [19]. The Adam algorithm is an adaptive gradient algorithm that adapts the value of the learning rate if the network does not improve the performance during training. We set the threshold of learning rate modification after 30 epochs, and it decayed $1e-6$. The initial learning rate was set to $1e-4$.

The initialization weights' algorithm used was Kaiming [19,20], and the loss function used was the dice coefficient for multiclass segmentation.

Figure 3. Illustration of the additive attention gate [15]. Reproduced from the cited source which is published under Creative Commons Attribution 4.0 International License [21].

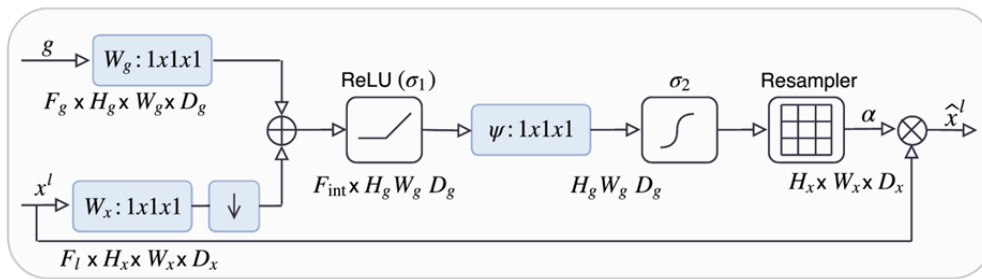
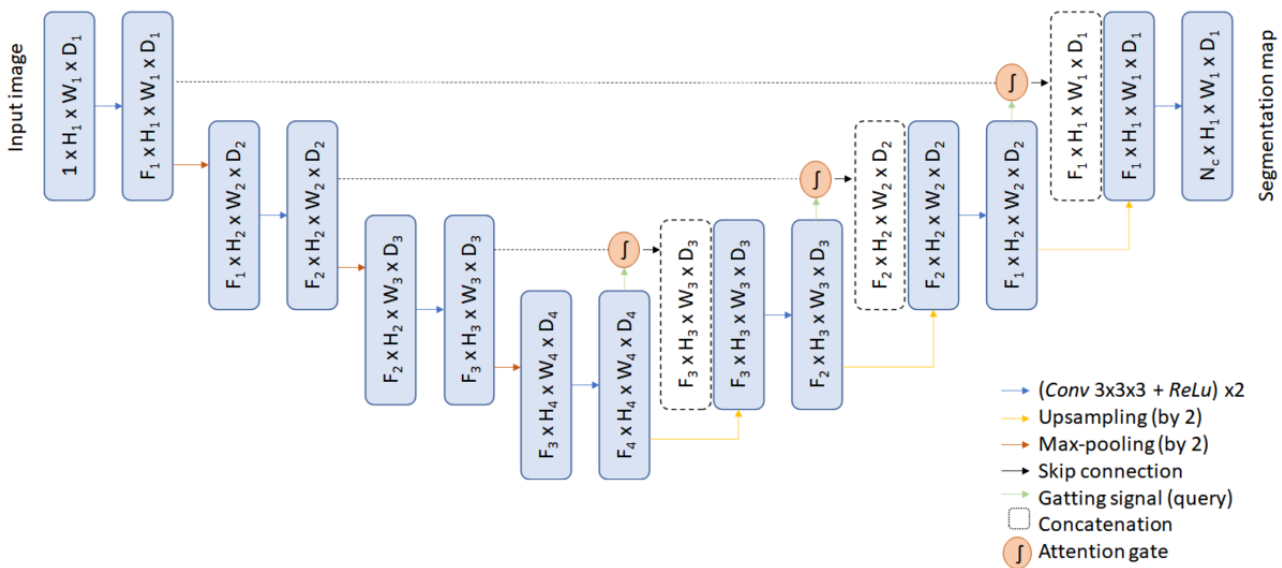


Figure 4. Scheme of the deep neural network architecture [15]. Fl: Feature map in the layer l; H: height; W: width; D: dimension; Conv 3x3: convolution operation with a 3x3 kernel; ReLu: rectified linear unit operation. Reproduced from the cited source which is published under Creative Commons Attribution 4.0 International License [21].



Results

The goal of this work was to implement a pipeline for PCLs detection on CT scan images as well as the pancreas. This was performed with a two-step pipeline formed by a first preprocessing that consisted of a normalization of all the data sets with a soft tissue normalization technique centered at an HU of 50. This value was selected since it is the state-of-the-art value assigned to the organs and it matches with the mean HU of the pancreas calculated for all the studies in our data set. Afterwards a central crop of the CT was applied; from a slice size of 512x512 to 240x240 after the central cropping to just focus on the center of the abdomen (anatomic location of the pancreas). Finally, the network was trained with random patches of 160x160 of this central crop, and therefore, the inference consisted of iterating around this central crop of multiple inferences of patches of 160x160.

During the inference, the test-time augmentation (TTA) technique was applied. For every CT, 4 geometrical transformations were used. Multiple options were considered in which way the TTA should be applied; however, we concluded that translation and rotation transformations were the most accurate since, for example, flipping would just confuse the network. Hence, after studying multiple options, a positive rotation of 7 degrees and a negative rotation of 11 degrees as

well as 2 positive translations of 5 and 10 pixels were considered. Positive and negative rotations were considered since in CT scans the abdomen can be tilted one way and the other, but higher values for both rotation and translation would just result in bad predictions. Using more TTA transformations were ruled out due to the latency that this adds to the final pipeline. The final result is a merging of this 4 TTA transformations inferred and the original CT without any transformation. We averaged the probability of each class, and after having them merged, a softmax function was applied for obtaining the final binarized image [22].

Finally, a postprocessing pipeline was implemented to improve the segmentation results performed by the network and minimize the number of false positive detections. First, a mask of the abdomen was generated and eroded to eliminate wrong predictions in the edges of the abdomen, where the pancreas anatomically is not found. Secondly, all segmented cysts that were not in touch with the pancreas were also removed. Finally, we established a minimum of 10 voxels to consider a predicted cyst as true positive. Therefore, if there were some randomly segmented pixels considered as cysts that were not previously filtered, they were ignored. Images with qualitative results of this method are shown in Figure 5.

The fully automated segmentation was performed on a modern computer with an NVIDIA GPU T4 to automatically detect

PCLs in abdominal CT scans. The programming language used was Python and the framework for the model development was PyTorch. The sensitivity for all cases was 93.1% (SD 0.1%), and the specificity was 81.8% (SD 0.1%).

Additionally, due to the small amount of some subtypes of pancreatic cysts in the training database (Figure 6), we considered it reasonable to divide the whole cohort of patients into 2 big groups: on the one hand, the most dangerous cyst types, bearing malignant potential (IPMN and MCN), and on the other hand, the ones with malignant potential close to 0 (PCYST and SCA). If we consider this classification, the global specificity and sensitivity for the detection of the most dangerous group were 81.8% and 97.0%, respectively, while for the least dangerous ones, they were 81.8% and 89.0%, respectively.

One of the main metrics used to evaluate the effectiveness of this method was the sensitivity or true positive rate. This is something to highlight since it is better to have a false positive than a false negative in this study due to the consequences of obtaining each one: for a false positive, a review of the detection would be needed, but for a false negative, the consequences are much worse because a PCL can exist and not be detected. If we compare the most dangerous group and the least dangerous group, meaning the one that can easily evolve to pancreatic cancer versus the one that cannot evolve to pancreatic cancer as easily, it is a remarkable fact that the sensitivity is almost 10% higher for the dangerous group, which makes the network even more efficient. Having a better true positive rate for the most dangerous group rather than for the least dangerous group is a highlight of this study.

Figure 5. Illustration of the qualitative results obtained. Each pair of images belongs to a patient with a pancreatic cystic lesion. The left image of the pair is the ground truth, while the right one is the outcome of this method. The pixels that belong to the pancreas are painted in green and the ones for the pancreatic cystic lesion in red.

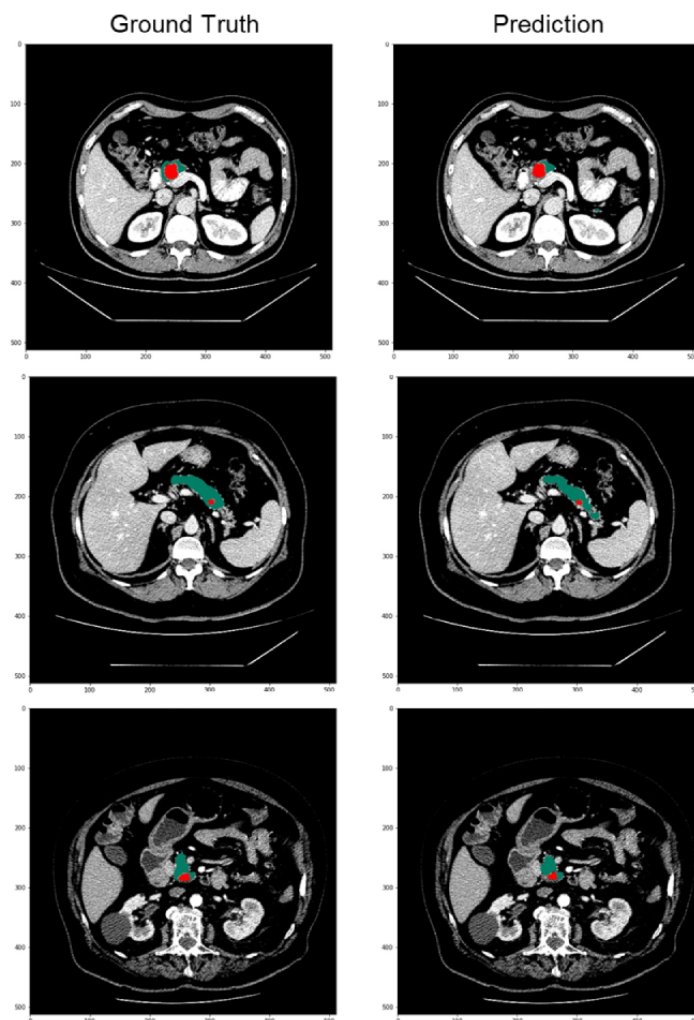
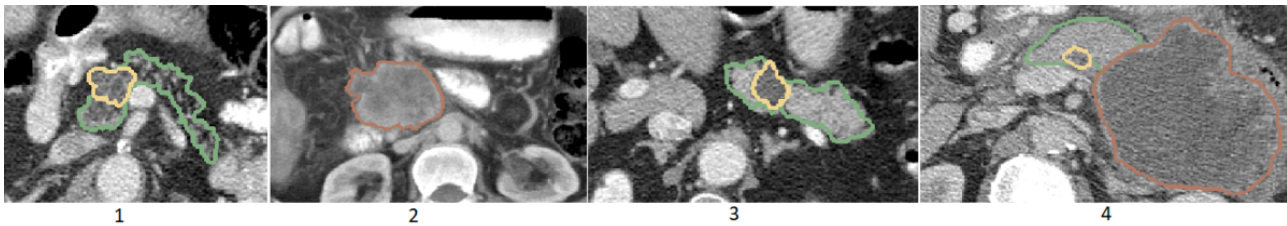


Figure 6. Example of the types of pancreatic cysts included in this research. (1) Serous cystadenoma, marked in yellow. (2) Mucinous cystic neoplasm, marked in red. (3) Intraductal papillary mucinous neoplasm, marked in yellow. (4) Pseudocyst, marked in red). Pancreas is depicted in green.



Discussion

Principal Findings

In this study, we applied and validated an AGNet deep neural network to detect PCLs. The aim was to assist imaging specialists for a better diagnosis, and therefore, achieve better determining of treatment plans. First, a pancreatic CT image database with different types of cyst present was created based on the diagnosis of anatomical pathology or an imaging specialist. From this database, we established an AI system for the automatic detection of pancreatic cysts (with further classification) and then validated it in a test experiment.

In our study, the sensitivity for the detection of PCL was 93.1% (SD 0.1%), and the specificity was 81.8% (SD 0.1%), demonstrating that PCLs can be automatically detected by AI with a diagnostic performance comparable to radiologists.

This is significant because even though AI has shown excellent performance for segmentation of organs with sharp borders, in organs with vague delineation like the pancreas (eg, caused by fat interdigitations), the detection of lesions remains a difficult task for algorithms [23].

In a previous work (Abel et al [14]), an overall sensitivity of 78.8% for the detection of pancreatic cysts was obtained. The maximum sensitivity was seen in big lesions, ranging from 87.8% for cysts under 220 mm³ to 96.2% for tumors in the distal pancreas. Importantly, in this work, they analyzed the size of the lesion by volume, and in our study, we analyzed it with the diameter of the biggest slice of the lesion. Another difference between this work and ours is the deep learning architecture they used. They used an nnUNet pretrained, and we used an attention gate U-Net without pretraining.

Overall, these results demonstrate that an automated detection of PCL on CT scans is feasible.

Nevertheless, limitations to our research are still present. Although the results obtained indicate that the diagnostic accuracy is comparable to that of radiologists, it is important to bear in mind that this research intends to develop an assistive tool, not to be in any case a substitute for doctors. Moreover, this is a retrospective single-center analysis study. To further evaluate and validate the clinical applicability, next steps would include a prospective study on multicenter clinical data.

Importantly, the possibility for malignancy varies across various forms of PCLs. Therefore, precise cyst characterization is crucial for proper care. The most clinically significant distinction is separating nonmucinous cystic lesions from mucinous cystic lesions, which have malignant potential and may benefit from surgical removal. However, distinction between cyst types is difficult in a clinical setting.

Due to the lack of data for each specific subtype of PCL, this study only aimed at detecting but not classifying PCLs. Next steps would include increasing the final data set size to further assess and validate the classification performance of a deep neural network, which would have a significant effect in clinical practice.

Limitations

PCL detection algorithm was trained and tested on data from a single hospital, which limited the available amount of data and hindered the possibility to perform an external validation.

As previously mentioned, the data in the training database were divided into 2 big groups (IPMN and MCN vs pseudocysts and SCA) due to the lack of data for each specific subtype of pancreatic cysts. For further validation, not only detection but also classification, more data are needed for the training database for each of the cyst subtypes that we are willing to differentiate.

Next steps will be to obtain images from other hardware manufacturers and improve our database. This will need to be studied thoroughly to make the images from different hospitals compatible to each other. Another approach to improve the data set is to widen the samples of each type of cyst to make it more heterogeneous.

Conclusions

This study presents a clinical validation for automated detection of PCLs using an AGNet deep neural network. Based on the validation of an artificial deep neural network [15], results indicate that AI can be a feasible tool to help radiologist to cope with the increasing demand of cross-sectional imaging tests. The proposed method shows ability to obtain an accurate diagnosis. This artificial network, working together with specialists, proves to be a potential and effective way to tackle the early detection of pancreatic cancer.

Authors' Contributions

MMD, NTF, and MTFP (Hospital de Mataró, Consorci Sanitari del Maresme, Barcelona, Spain) were responsible for data collection, anonymization, experiment design, and result validation. MRM, DC, JRC, and JGL performed the experiments. All authors contributed to the writing, revision, and final approval of the manuscript.

Conflicts of Interest

JGL and JRC are founders of Sycal Technologies and declare significant ownership. MRM and DC are employed by Sycal Technologies. The other authors report no conflicts of interest.

References

1. Ilic M, Ilic I. Epidemiology of pancreatic cancer. *World J Gastroenterol* 2016 Nov 28;22(44):9694-9705 [FREE Full text] [doi: [10.3748/wjg.v22.i44.9694](https://doi.org/10.3748/wjg.v22.i44.9694)] [Medline: [27956793](https://pubmed.ncbi.nlm.nih.gov/27956793/)]
2. Branca Vergano L, Monesi M, Vicenti G, Bizzoca D, Solarino G, Moretti B. Posterior approaches in malleolar fracture: when, why and how. *J Biol Regul Homeost Agents* 2020;34(3 Suppl. 2):89-95. [Medline: [32856446](https://pubmed.ncbi.nlm.nih.gov/32856446/)]
3. Mizrahi JD, Surana R, Valle JW, Shroff RT. Pancreatic cancer. *Lancet* 2020 Jun 27;395(10242):2008-2020. [doi: [10.1016/S0140-6736\(20\)30974-0](https://doi.org/10.1016/S0140-6736(20)30974-0)] [Medline: [32593337](https://pubmed.ncbi.nlm.nih.gov/32593337/)]
4. Chernyak V, Flusberg M, Haramati LB, Rozenblit AM, Bellin E. Incidental pancreatic cystic lesions: is there a relationship with the development of pancreatic adenocarcinoma and all-cause mortality? *Radiology* 2015 Jan;274(1):161-169 [FREE Full text] [doi: [10.1148/radiol.14140796](https://doi.org/10.1148/radiol.14140796)] [Medline: [25117591](https://pubmed.ncbi.nlm.nih.gov/25117591/)]
5. Laffan TA, Horton KM, Klein AP, Berlanstein B, Siegelman SS, Kawamoto S, et al. Prevalence of unsuspected pancreatic cysts on MDCT. *AJR Am J Roentgenol* 2008 Sep;191(3):802-807 [FREE Full text] [doi: [10.2214/AJR.07.3340](https://doi.org/10.2214/AJR.07.3340)] [Medline: [18716113](https://pubmed.ncbi.nlm.nih.gov/18716113/)]
6. Ip IK, Mortelet KJ, Prevedello LM, Khorasani R. Focal cystic pancreatic lesions: assessing variation in radiologists' management recommendations. *Radiology* 2011 Apr;259(1):136-141. [doi: [10.1148/radiol.10100970](https://doi.org/10.1148/radiol.10100970)] [Medline: [21292867](https://pubmed.ncbi.nlm.nih.gov/21292867/)]
7. Lee KS, Sekhar A, Rofsky NM, Pedrosa I. Prevalence of incidental pancreatic cysts in the adult population on MR imaging. *Am J Gastroenterol* 2010 Sep;105(9):2079-2084. [doi: [10.1038/ajg.2010.122](https://doi.org/10.1038/ajg.2010.122)] [Medline: [20354507](https://pubmed.ncbi.nlm.nih.gov/20354507/)]
8. Zhang X, Mitchell DG, Dohke M, Holland GA, Parker L. Pancreatic cysts: depiction on single-shot fast spin-echo MR images. *Radiology* 2002 May;223(2):547-553. [doi: [10.1148/radiol.2232010815](https://doi.org/10.1148/radiol.2232010815)] [Medline: [11997566](https://pubmed.ncbi.nlm.nih.gov/11997566/)]
9. Elta GH, Enestvedt BK, Sauer BG, Lennon AM. ACG clinical guideline: diagnosis and management of pancreatic cysts. *Am J Gastroenterol* 2018 Apr;113(4):464-479. [doi: [10.1038/ajg.2018.14](https://doi.org/10.1038/ajg.2018.14)] [Medline: [29485131](https://pubmed.ncbi.nlm.nih.gov/29485131/)]
10. European Study Group on Cystic Tumours of the Pancreas. European evidence-based guidelines on pancreatic cystic neoplasms. *Gut* 2018 May;67(5):789-804 [FREE Full text] [doi: [10.1136/gutjnl-2018-316027](https://doi.org/10.1136/gutjnl-2018-316027)] [Medline: [29574408](https://pubmed.ncbi.nlm.nih.gov/29574408/)]
11. Springer S, Masica DL, Dal Molin M, Douville C, Thoburn CJ, Afsari B, et al. A multimodality test to guide the management of patients with a pancreatic cyst. *Sci Transl Med* 2019 Jul 17;11(501):eaav4772 [FREE Full text] [doi: [10.1126/scitranslmed.aav4772](https://doi.org/10.1126/scitranslmed.aav4772)] [Medline: [31316009](https://pubmed.ncbi.nlm.nih.gov/31316009/)]
12. Cai J, Lu L, Zhang Z, Xing F, Yang L, Yin Q. Pancreas segmentation in MRI using graph-based decision fusion on convolutional neural networks. *Med Image Comput Comput Assist Interv* 2016:442-450. [doi: [10.1007/978-3-319-46723-8_51](https://doi.org/10.1007/978-3-319-46723-8_51)] [Medline: [28083570](https://pubmed.ncbi.nlm.nih.gov/28083570/)]
13. Hosny A, Parmar C, Quackenbush J, Schwartz LH, Aerts HJWL. Artificial intelligence in radiology. *Nat Rev Cancer* 2018 Dec;18(8):500-510 [FREE Full text] [doi: [10.1038/s41568-018-0016-5](https://doi.org/10.1038/s41568-018-0016-5)] [Medline: [29777175](https://pubmed.ncbi.nlm.nih.gov/29777175/)]
14. Abel L, Wasserthal J, Weikert T, Sauter AW, Nestic I, Obradovic M, et al. Automated detection of pancreatic cystic lesions on CT using deep learning. *Diagnostics (Basel)* 2021 May 19;11(5):901 [FREE Full text] [doi: [10.3390/diagnostics11050901](https://doi.org/10.3390/diagnostics11050901)] [Medline: [34069328](https://pubmed.ncbi.nlm.nih.gov/34069328/)]
15. Oktay O, Schlemper J, Folgoc LL, Lee M, Heinrich M, Misawa K, et al. Attention U-Net: learning where to look for the pancreas. *ArXiv Preprint* posted online April 11, 2018. [doi: [10.48550/arxiv.1804.03999](https://doi.org/10.48550/arxiv.1804.03999)]
16. Fedorov A, Beichel R, Kalpathy-Cramer J, Finet J, Fillion-Robin J, Pujol S, et al. 3D Slicer as an image computing platform for the Quantitative Imaging Network. *Magn Reson Imaging* 2012 Nov;30(9):1323-1341 [FREE Full text] [doi: [10.1016/j.mri.2012.05.001](https://doi.org/10.1016/j.mri.2012.05.001)] [Medline: [22770690](https://pubmed.ncbi.nlm.nih.gov/22770690/)]
17. Huo Y, Tang Y, Chen Y, Gao D, Han S, Bao S, et al. Stochastic tissue window normalization of deep learning on computed tomography. *J Med Imag* 2019 Oct 1;6(04):1. [doi: [10.1117/1.jmi.6.4.044005](https://doi.org/10.1117/1.jmi.6.4.044005)]
18. Weng W, Zhu X. INet: convolutional networks for biomedical image segmentation. *IEEE Access* 2021;9:16591-16603. [doi: [10.1109/access.2021.3053408](https://doi.org/10.1109/access.2021.3053408)]
19. Kingma D, Ba J. Adam: a method for stochastic optimization. *ArXiv Preprint* posted online Dec 22, 2014. [doi: [10.48550/arxiv.1412.6980](https://doi.org/10.48550/arxiv.1412.6980)]
20. Cornell RB, Nissley SM, Horwitz AF. Cholesterol availability modulates myoblast fusion. *J Cell Biol* 1980 Sep;86(3):820-824 [FREE Full text] [doi: [10.1083/jcb.86.3.820](https://doi.org/10.1083/jcb.86.3.820)] [Medline: [7410480](https://pubmed.ncbi.nlm.nih.gov/7410480/)]
21. Attribution 4.0 International (CC BY 4.0). Creative Commons. URL: <https://creativecommons.org/licenses/by/4.0/> [accessed [2023-03-16]]

22. Seçkin AM, Berens P. Test-time Data Augmentation for Estimation of Heteroscedastic Aleatoric Uncertainty in Deep Neural Networks. *Test-time Data Augmentation for Estimation of Heteroscedastic Aleatoric Uncertainty in Deep Neural Networks* 2022 Jul 11:1-9.
23. Anta JA, Martínez-Ballesteros I, Eiroa D, García J, Rodríguez-Comas J. Artificial intelligence for the detection of pancreatic lesions. *Int J Comput Assist Radiol Surg* 2022 Oct 11;17(10):1855-1865. [doi: [10.1007/s11548-022-02706-z](https://doi.org/10.1007/s11548-022-02706-z)] [Medline: [35951286](https://pubmed.ncbi.nlm.nih.gov/35951286/)]

Abbreviations

AG: attention gate

AI: artificial intelligence

CT: computed tomography

IPMN: intraductal pseudopapillary mucinous neoplasia

HU: Hounsfield unit

MCN: mucinous cystic neoplasm

PCL: pancreatic cystic lesion

SCA: serous cystadenoma

TTA: test-time augmentation

Edited by K El Emam; submitted 01.07.22; peer-reviewed by F Maleki, W Klement; comments to author 17.08.22; revised version received 02.09.22; accepted 11.11.22; published 17.03.23.

Please cite as:

Duh MM, Torra-Ferrer N, Riera-Marín M, Cumelles D, Rodríguez-Comas J, García López J, Fernández Planas MT

Deep Learning to Detect Pancreatic Cystic Lesions on Abdominal Computed Tomography Scans: Development and Validation Study
JMIR AI 2023;2:e40702

URL: <https://ai.jmir.org/2023/1/e40702>

doi: [10.2196/40702](https://doi.org/10.2196/40702)

PMID: [38875547](https://pubmed.ncbi.nlm.nih.gov/38875547/)

©Maria Montserrat Duh, Neus Torra-Ferrer, Meritxell Riera-Marín, Dídac Cumelles, Júlia Rodríguez-Comas, Javier García López, M^a Teresa Fernández Planas. Originally published in JMIR AI (<https://ai.jmir.org>), 17.03.2023. This is an open-access article distributed under the terms of the Creative Commons Attribution License (<https://creativecommons.org/licenses/by/4.0/>), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR AI, is properly cited. The complete bibliographic information, a link to the original publication on <https://www.ai.jmir.org/>, as well as this copyright and license information must be included.

Original Paper

Patient Embeddings From Diagnosis Codes for Health Care Prediction Tasks: Pat2Vec Machine Learning Framework

Edgar Steiger¹, DPhil; Lars Eric Kroll¹, DPhil

Zi Data Science Lab, Department IT and Data Science, Central Research Institute of Ambulatory Health Care in Germany (Zi), Berlin, Germany

Corresponding Author:

Edgar Steiger, DPhil

Zi Data Science Lab

Department IT and Data Science

Central Research Institute of Ambulatory Health Care in Germany (Zi)

Salzufer 8

Berlin, 10587

Germany

Phone: 49 40052485

Email: esteiger@zi.de

Abstract

Background: In health care, diagnosis codes in claims data and electronic health records (EHRs) play an important role in data-driven decision making. Any analysis that uses a patient's diagnosis codes to predict future outcomes or describe morbidity requires a numerical representation of this diagnosis profile made up of string-based diagnosis codes. These numerical representations are especially important for machine learning models. Most commonly, binary-encoded representations have been used, usually for a subset of diagnoses. In real-world health care applications, several issues arise: patient profiles show high variability even when the underlying diseases are the same, they may have gaps and not contain all available information, and a large number of appropriate diagnoses must be considered.

Objective: We herein present Pat2Vec, a self-supervised machine learning framework inspired by neural network-based natural language processing that embeds complete diagnosis profiles into a small real-valued numerical vector.

Methods: Based on German outpatient claims data with diagnosis codes according to the International Statistical Classification of Diseases and Related Health Problems, 10th Revision (ICD-10), we discovered an optimal vectorization embedding model for patient diagnosis profiles with Bayesian optimization for the hyperparameters. The calibration process ensured a robust embedding model for health care-relevant tasks by aggregating the metrics of different regression and classification tasks using different machine learning algorithms (linear and logistic regression as well as gradient-boosted trees). The models were tested against a baseline model that binary encodes the most common diagnoses. The study used diagnosis profiles and supplementary data from more than 10 million patients from 2016 to 2019 and was based on the largest German ambulatory claims data set. To describe subpopulations in health care, we identified clusters (via density-based clustering) and visualized patient vectors in 2D (via dimensionality reduction with uniform manifold approximation). Furthermore, we applied our vectorization model to predict prospective drug prescription costs based on patients' diagnoses.

Results: Our final models outperform the baseline model (binary encoding) with equal dimensions. They are more robust to missing data and show large performance gains, particularly in lower dimensions, demonstrating the embedding model's compression of nonlinear information. In the future, other sources of health care data can be integrated into the current diagnosis-based framework. Other researchers can apply our publicly shared embedding model to their own diagnosis data.

Conclusions: We envision a wide range of applications for Pat2Vec that will improve health care quality, including personalized prevention and signal detection in patient surveillance as well as health care resource planning based on subcohorts identified by our data-driven machine learning framework.

(JMIR AI 2023;2:e40755) doi:[10.2196/40755](https://doi.org/10.2196/40755)

KEYWORDS

electronic health records; ICD; machine learning; health care; data; diagnosis; model; drug; drug prescription; performance; applications; quality; prevention

Introduction

Public health surveillance and health care research in many countries depend on electronic health records (EHRs), including claims data [1-4]. In these records, patients' medical diagnoses are often coded according to a string-based disease classification convention, for example, the International Statistical Classification of Diseases and Related Health Problems, 10th Revision (ICD-10) [5]. Their sequence of ICD codes characterizes the medical history of every patient.

Common tasks in clinical, epidemiological, or health care research on claims data expect numerical input (eg, regression and classification tasks such as linear or logistic regression or advanced machine learning tools such as gradient-boosted trees and deep learning). These methods are often used to predict specific health outcomes [6-17] or the utilization of health care institutions [18-22].

To derive numerical input for these methods from the string-based diagnosis profiles, a procedure called binary encoding (or binarization, one-hot encoding) is most often used [6-11,15-17,20-24]. Using binary encoding, diagnoses are represented numerically by either 1 or 0, if the patient had or did not have the chosen diagnosis, respectively. As the pool of possible diagnostic codes is vast, binary encoding usually relies on a selected subset of diagnoses chosen by either field experts [6,16] or data-driven feature selection [10,15,17]. Diagnoses can also be represented by the number of times they appear [9,12,25,26]. Most often, they are pooled into clinical groups before further analysis [18-22,24,27-29].

Ideally, a disease classification such as ICD-10 would only cover clearly distinguishable medical conditions and concepts, but in reality, we have to deal with overlaps and uncertainties. Therefore, a faithful numerical representation of the patient's medical history needs to take into account that different ICD codes may represent similar or even identical underlying issues. Frequently, computational and methodological constraints limit the number of diagnoses and interaction effects that can be considered. Binary encoding suffers in this regard, as it considers medical diagnoses as distinctive and unrelated features. As such, it limits the methodical progress of prediction tasks on claims data, especially the application of advanced machine learning methods. Thus, other methods of numerical representation of ICD diagnosis codes should be investigated to enable better individual health care and more precise prediction of health care demand.

We investigate herein how a real-valued numerical representation (or vectorization, embedding) (see Chapter 15 in [30]) of patients' medical diagnosis profiles that uses their whole diagnostic ICD profiles can be derived. This embedding should compress the information from up to 14,877 possible 5-digit International Statistical Classification of Diseases and Related Health Problems, 10th revision, German Modification (ICD-10-GM) 2019 [31] codes, improve the performance of common health care prediction tasks, and let advanced (nonlinear) machine learning methods reach their full potential when used on claims data.

To find such an embedding, we employ a self-supervised machine learning algorithm inspired by natural language processing (NLP), namely, Doc2Vec [32], which itself is an extension of Word2Vec [33,34]. It has been applied to nonlanguage-specific tasks before [35-37]. Many studies [14,29,38-42] have investigated embeddings of the ICD codes themselves, whereas some [14,25,42] arrived at patient-level embeddings for specific prediction tasks (Supplementary Table S1 in [Multimedia Appendix 1](#)). Here, we want to broaden the scope of the possible applications to general health care-related questions. It has been shown that hyperparameter tuning for Word2Vec and Doc2Vec can lead to considerably better results, especially on nonlanguage-related tasks [35,37]. As such, we employ a Bayesian search on a hyperparameter grid to identify an optimal model for the vector embedding procedure. We evaluate our embedding model on broad health care prediction tasks with standard (linear and logistic regression) and advanced machine learning techniques (gradient-boosted trees). We also test how well the vectorization works with smaller data sets and how well it handles missing data with random data dropout sampling. In addition, we inspect the results visually in a 2D projected space along with a clustering of the embedded patient profiles to reveal the properties of our cohort. Finally, we evaluate the resulting vectorization model for the health care-relevant task of predicting drug spending at the patient level.

Our method gave better results than binary encoding, but only after tuning the hyperparameters and on large enough data sets. The compression of the information of thousands of ICD-10 codes into a vector space of no more than 100 dimensions was achieved. We observed large performance gains using gradient-boosted trees with the vector embedding over classic linear or logistic regression with binary-encoded data. In addition, the vectorization models are more robust to missing data than baseline binary encoding. The final model learned on our extensive data can be shared and used by other stakeholders on much smaller data sets (eg, for supervised machine learning methods that predict clinical or other health care outcomes).

Methods

Data

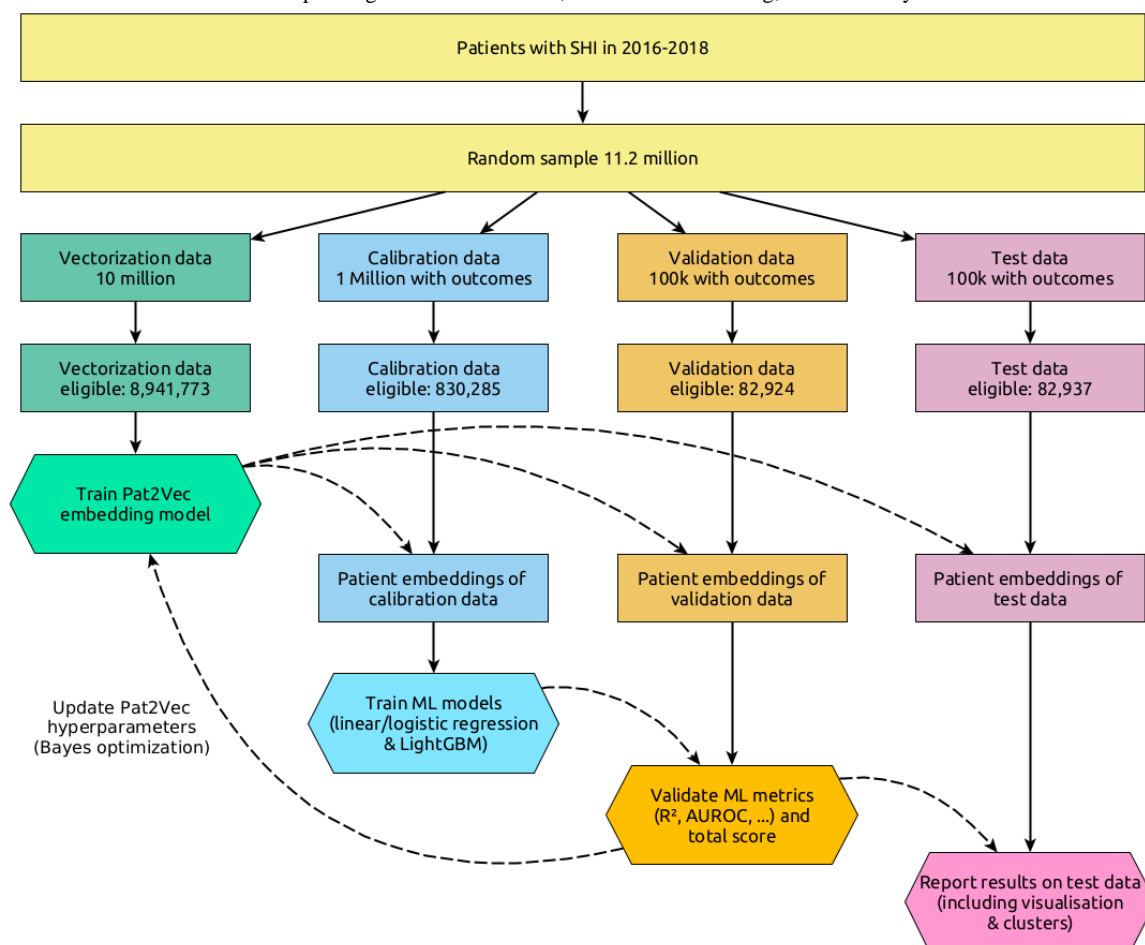
The diagnosis data are based on comprehensive nationwide outpatient claims data from 2016 to 2019 of all patients with statutory health insurance (SHI) in Germany. According to the Federal Statistical Office [43], there were 73,009,237 persons eligible for the SHI (87.8% of the population) in 2019. The pseudonymous data include diagnoses for all people in Germany with SHI who visited an outpatient physician in 2016 or later. Among others, the data include demographic characteristics such as age and gender, as well as diagnoses with markers of certainty and other billing-relevant information. These data do not contain information on inpatient treatment in hospitals. Diagnoses are coded according to the ICD-10-GM [31]. In addition to the diagnosis data, we extracted individual information on prescribed and dispensed medications from the pseudonymous data of nationwide outpatient drug prescriptions.

The claims data and the prescription data are linked by patient information (compare [44]).

We chose $N=11,200,000$ patients at random from the full population of people with SHI because technical limitations make it impossible to use the full data. To achieve this study sample size, we shuffled all patients in the claims database randomly and selected the top N records for the sample. All patients with at least one data entry after 2016 were eligible. The sample is divided into 4 data sets by random subsampling from the study population (Textbox 1).

These samples were filtered for patients with consistent information regarding gender and age during the years considered for analysis (2016 to 2019). The training data in (1) for the vectorization model were restricted to ICD-10 codes (5-digit notation) from 2016 to 2018, whereas the calibration, validation, and test sets in (2)-(4) were restricted to codes from 2018. Only patients with at least one confirmed diagnosis during the period in question were kept. This left us with sample sizes of 8,941,773 (vectorization training), 830,285 (calibration training), 82,924 (validation), and 82,937 (test), see Figure 1.

Figure 1. Flowchart of data sampling and algorithmic schematic. Patient data flows are represented by solid, straight lines, while machine learning models and other meta-information flows are represented by dashed, curved lines. Rectangles are patient data, while hexagons are algorithms or analysis methods. AUROC: area under the receiver operating characteristic curve; ML: machine learning; SHI: statutory health insurance.



Because of the regulations of the German health care system (see “The German Health Care System” in [45], or a more detailed description of the German system in [46]), diagnoses are available on a quarterly basis (but without temporal order within a quarter), with reference to cases and places of treatment. As such, we generated a sequence of codes for each patient with a certain temporal order: confirmed diagnoses are grouped by case and place of treatment, and these groups are ordered by temporal succession of quarters, but if more than 1 group appears within one-quarter, these groups are shuffled randomly within the quarter (as well as diagnoses within a group).

Furthermore, when training the model (see below), only diagnoses that were seen at least 100 times in the training data were taken into account.

As health care-relevant outcomes in (2)-(4), we used 4 different quantities for calibration: the *number of cases* (a proxy indicator for the number of medical consultations), (ambulatory) *emergency health care utilization*, *age*, and *gender*. The number of cases in 2019 is approximate due to data limitations: a case is defined as the unique combination of a quarter, a patient, a treating medical facility, the billing association of SHI physicians, and the time stamp of data processing. The binary outcome of emergency health care utilization is 1 if at least one case in 2019 of the respective patient was billed as an emergency, and 0 otherwise. The sociodemographic variables age (in years) and gender (binary-encoded) were also extracted from the data.

As data for robustness analysis against diagnosis dropout, we randomly dropped 10%, 25%, or 50% of diagnosis codes for each patient (rounded to nearest number, but kept at least one code).

As data for robustness analysis against varying training data set sizes, we used different percentages of the original vectorization

training data (reducing the vectorization data from 10 million patients to 10,000 patients).

For a further analysis, we extracted the drug prescription costs from the ambulatory drug prescription data of residents in Germany with SHI. These costs are the total (in euros) of all billed prescribed drugs for the respective patient in 2019 (if any, otherwise 0).

Textbox 1. Data sets obtained by random subsampling from the study population.

<p>1. Vectorization</p> <p>A total of 10,000,000 patients as a vectorization training set for self-supervised machine learning to learn a model for numerical representation (embedding) of patients' profiles.</p> <p>2. Calibration</p> <p>A total of 1,000,000 patients with embeddings based on a model from (1) serving as a calibration training set for supervised machine learning on prediction tasks.</p> <p>3. Validation</p> <p>A total of 100,000 patients with embeddings based on a model from (1) serving as a validation set for the calibration prediction models learned in (2) and, in turn, hyperparameter tuning of vectorization in (1).</p> <p>4. Test</p> <p>A total of 100,000 patients as a test set for final analysis and presentation of the results.</p>

Ethical Considerations

The use of claims data for this analysis is governed by the German Code of Social Law (SGB X 80 in conjunction with SGB V 68c): our study aims to improve health care quality by exploring diagnoses profiles and predicting health care-relevant outcomes. While approval and consent of individual human patients within the cohort are operationally impossible to acquire, they are also not required by the German Code of Social Law as we used deidentified, routinely collected data in a retrospective study. In addition, we argue that the conclusions we can draw from our analyses are in the best interest of patients and will improve future public health services.

Binary Encoding and Baseline Model

Binary encoding creates a data matrix with rows for patients and columns for variables. Each variable represents one of the diagnoses being looked at (out of a chosen subset of all available diagnoses) and is given a 1 in the corresponding row and column if the patient had that diagnosis and a 0 if they did not.

Here, we employ such a binary encoding approach as a baseline model: First, we sorted all confirmed unique ICD-10 diagnosis codes from 2019 by the number of patients with this diagnosis in the data. Second, for a given number M of top diagnoses and the sample patients from above, we formed the appropriate data matrix with M columns corresponding to the top M diagnoses and each row representing a patient, using binary encoding like described above. This is the baseline model for numerization of the diagnosis codes and will be compared with the real-valued patient-level embedding described in the next section.

ICD2Vec and Pat2Vec

Similar to [14], we used an advanced approach to a real-valued embedding of diagnosis codes, applying a method from NLP called Word2Vec and its extension Doc2Vec [32-34]. Trained

on a corpus of text data, Word2Vec vectorizes individual words and keeps their semantic meaning by mapping similar or related words to similar vectors (according to multidimensional distance measures in a Euclidean space) and antagonistic words to diverging vectors. As an extension to Word2Vec, the Doc2Vec algorithm also learns vectors for each document. Similar documents are represented by vectors that are similar to those of the similar documents.

Word2Vec is in fact a (shallow) neural network in the sense that individual words are represented by vectors (embeddings) of a fixed size, and the entries of these vectors are used directly to predict the vectors of other words in a single-layer neural network; that is, the embeddings are themselves the parameters of the single hidden layer. Word2Vec goes over every word in each document step-by-step and repeatedly during training and updates the neural network's parameters (or rather, the embeddings) by either predicting from the current word the neighboring or context words as targets (skip-gram) or predicting a target word from the neighboring or context words (continuous bag of words) [33]. In both cases, the update to the network's parameters after training on a single word would include updating all parameters for all words that are not in the context. For computational efficiency (because of large vocabularies), this is circumvented by either updating only some negative examples of words that are not in the context of the word under consideration [34] or by applying a hierarchical softmax to the network update [33]. In fact, it is also possible to apply both techniques at the same time.

Doc2Vec is an extension to the Word2Vec algorithm in the sense that it is applied in parallel to Word2Vec. Additionally, while learning the vector embeddings of every word in the corpus, the vector embeddings of the documents that form the corpus are learned in the same manner. Doc2Vec can be trained in 2 different ways [32]: either with "distributed memory" (DM);

similar to Word2Vec's continuous bag of words), where each target word from the document is predicted using both the context words and the document's embedding, or with "distributed bag of words" (DBOW; similar to Word2Vec's skip-gram), where target words from the document are predicted using the document itself and separately updating the context words.

For more background on neural networks and how they are applied to NLP tasks, see [47] and [48].

In our framework, we treat every ICD-10 diagnosis code as a word and the sequence of diagnosis codes for a patient as a document. These documents are our corpus data for training ICD2Vec (by applying Word2Vec to ICD-10 codes) and Pat2Vec (by applying Doc2Vec to patients' sequences of diagnosis codes).

For training the 2Vec algorithms, we have to choose a vector size of M (among other parameters; see below). Pat2Vec is trained on the patients' sample data and then gives us a data

matrix with M columns, where each row or patient is a vector of length M (the embedding of the corresponding patient), encoding *all* of their diagnoses. Additionally, we obtain in parallel a vectorization of the ICD-10 codes themselves (Word2Vec/ICD2Vec), where each code is represented by a vector.

Hyperparameter Tuning

The 2Vec algorithms need several parameters as input for the training of the vectorization model. These are referred to as hyperparameters and have different considered ranges (Textbox 2).

Following previous research [35,37], we tuned the hyperparameters for the vectorization model using a Bayesian hyperparameter optimization [49] over the ranges given above. We calibrated and validated the resulting vectorization models with supervised machine learning (see the next section) using the holdout calibration and validation data on the 4 calibration outcomes.

Textbox 2. Hyperparameters and their ranges.

1.	Vector size (100)	Length of the vector assigned to each patient. We hold this fixed while tuning the hyperparameters, but we will vary this value afterward for comparisons.
2.	Minimal count (100)	Only diagnoses that appear at least 100 times in the data are considered for anonymization purposes because of rare diseases. We will not optimize this parameter.
3.	Window size (1-10)	Describes how many of the neighboring codes will be considered in each training step within the 2Vec algorithm and a given sequence of codes.
4.	Downsampling	Smaller values of the downsampling parameter mean that more of the most common words will be randomly excluded from the training data (default 0.001). After preliminary analysis, we observed that downsampling is always detrimental to our task, so we did not downsample our data.
5.	Epochs (1-20)	The number of training epochs describes how many times each patient's code sequence will be looked at to update the vectorization model.
6.	Negative sampling (0-20)	For each update of a word and its neighboring words (within the window size range), this gives the number of random words not within the window that will be updated as negative examples; 0 for no negative sampling.
7.	Negative sampling exponent (-5 to 5)	Smoothing exponent for the updates of the negative samples.
8.	Hierarchical softmax (Boolean)	This parameter describes how the network parameters will be updated at the end of each training step; true for hierarchical softmax and false for no hierarchical softmax.
9.	Distributed memory or distributed bag of words (Boolean)	Training of document vectors in either distributed memory (DM) or distributed bag of words (DBOW) fashion (see above); true for DM and false for DBOW.
10.	Alpha (0.001-0.1)	Learning rate of the neural network updates.

Regression and Classification Methods

Overview

The data matrices generated by binary encoding or Pat2Vec served as input data for prediction algorithms on the 4 calibration outcomes (number of cases, emergency health care utilization, age, and gender). The employed algorithms are described below, where LightGBM refers to the light gradient-boosted machine algorithm [50].

Regression

For the real-valued count outcomes of age and number of cases, we employed 2 different regression techniques: linear regression and an ensemble decision tree-based regression algorithm with gradient boosting (LightGBM Regressor) [50-52]. We chose LightGBM over other gradient-boosted tree methods because of its performance and fast training time [50,53,54]. Linear regression does not have additional input parameters; LightGBM was used out of the box without parameter optimization. The goodness of fit was measured by the R^2 and 1 minus the relative mean absolute error (also known as Cumming predictive measure [CPM]) [55].

Classification

For the binary outcomes of gender and emergency usage, we employed 2 different classification techniques: logistic regression and an ensemble decision tree-based classification algorithm with gradient boosting (LightGBM Classifier) [50,52,56]. Logistic regression does not have additional input parameters; LightGBM was used out of the box without parameter optimization. The goodness of fit was measured by the area under the receiver operating characteristic curve and the area under the precision-recall curve.

Final Model

The final model was chosen with Bayesian optimization of the hyperparameters by aggregating the 16 performance measures: 2 approaches with linear/logistic regression and gradient-boosted trees, and 2 measures for each of the 4 outcomes (R^2 and CPM for regression, receiver operating characteristic curve and area under the precision-recall curve for classification). All of these measures are in the range of 0 and 1, with higher values indicating better performance but varying in size and range between the 4 different outcomes and measures. As such, we took the performance measure values of the top 100 diagnoses baseline model as reference values. For each trial in the Bayesian optimization and its respective vectorization model, we calculated the 16 performance measures and divided them by the respective reference value from the top 100 diagnoses baseline model. We then aggregated these rates by calculating their arithmetic mean as a total score (ie, this gives a reference score of 1 for the top 100 diagnoses baseline model). The final model was chosen based on the best total score after this aggregation (Figure 1).

We then trained embedding models with the same hyperparameter configuration as the final model, but with different vector sizes M . Likewise, we derived the binary encoding matrices of the top M diagnoses for varying sizes of M . These embedding and binarization models were compared

on the same prediction tasks described above on the holdout test data. The same procedures were replicated on the different data sets for robustness analysis (diagnosis dropout and reduced training data size, respectively).

Additionally, we conducted an exploratory and visual analysis of the vector embeddings from the Pat2Vec vectorization on the test data. To this end, we projected the 100D patient vector embeddings into 2 dimensions using the uniform manifold approximation and projection (UMAP) algorithm [57]. In addition, these projections were clustered using hierarchical density-based clustering (hierarchical density-based spatial clustering of applications with noise [HDBSCAN]) [58]. We assessed the general demographic and health care properties of the clusters and identified overexpressed ICD-10 codes within each cluster as the codes that have the largest positive difference in their share within the respective cluster compared with their share in the general population. As an explainability analysis, we analyzed how ICD-10 diagnosis codes are associated with specific dimensions of the vector embedding of size 100. To this end, we calculated correlations over all patients in the test data between a subset of 60 relevant ICD-10 diagnosis codes, binary encoded per patient, and the 100 vector dimensions.

Furthermore, we predicted drug spending costs using the final embedding model with a vector size of 100 and the baseline model. We compared the performance (R^2 , mean absolute error, and CPM), again with linear regression and the gradient-boosted trees algorithm for regression (LightGBM Regressor). We also added age and gender as additional predictors to these models. Here, we tuned the hyperparameters of the LightGBM method using Bayesian optimization to achieve its full potential.

Software

Analysis was conducted primarily in the Python programming language (Python Software Foundation) [59], with additional analyses in the R statistical programming language (The R Foundation) [60]. Pat2Vec was implemented using the Gensim package [61] for Python with hyperparameter tuning via the Optuna package [62]. Machine learning prediction tasks were conducted with scikit-learn (linear and logistic regression, [63]) and the LightGBM Python package [50], while 2D projection and clustering were based on the UMAP package [57] and the HDBSCAN package [58], respectively. Final visualizations were prepared in R with the ggplot2 package [64].

Results

Sample Characteristics

After filtering the original sample of 11,200,000 patients, the data were limited to 9,937,919 patients. The average age of the patients was 45.2 years; 54.60% (5,426,481/9,937,919) of the cohort were female. The average number of cases per patient in 2019 was 8.4. About 18.32% (1,820,736/9,937,919) of the cohort had at least one emergency in 2019. The average drug spending in 2019 was €632.1 (US \$683.4). The average number of diagnosis codes from 2016 to 2018 (relevant for the training data) was 67.6, whereas the average number of codes in 2018 only (relevant for prediction tasks) was 34.6. Variance was very high on the variable drug spending, with an SD of 4383.9 (Table

1). Furthermore, we observed a high number of patients with a 0 value in drug spending in 2019 (2,132,938/9,937,919, 21.46%, patients).

Table 1. Patients' data characteristics.

Characteristics	Values
Age (years), mean (SD)	45.2 (24.1)
Female gender, n/N (%)	5,426,481/9,937,919 (54.60)
Number of cases, mean (SD)	8.4 (6.7)
Emergency in 2019, n/N (%)	1,820,736/9,937,919 (18.32)
Drug cost (€ ^a), mean (SD)	632.1 (4383.9)
Number of codes from 2016-2018, mean (SD)	67.6 (92.4)
Number of codes in 2018, mean (SD)	34.6 (45.5)

^a€=US \$1.08 (as of March 27, 2023).

Top M Diagnosis Codes

The baseline model was constructed from a binary encoding of the top M diagnosis codes, for varying numbers of M. The most prevalent diagnosis code was I10.90 (hypertension; 2,591,336/9,937,919, 26.08%, patients), followed by J06.9 (unspecified acute upper respiratory infection) and Z12.9 (unspecified special screening for neoplasms used in the various German cancer screening programs [65]). Many patients have at least one of the top diagnoses (eg, 8,947,182/9,937,919, 90.03%, patients) have at least one of the most prevalent diagnoses). By contrast, over 2000 unique diagnosis codes make up the bulk of the diagnoses, with a share of over 90% of all diagnosis codes (317,316,756/343,751,225, 92.31%) in the data (Supplementary Table S2 in [Multimedia Appendix 1](#)).

Hyperparameter Tuning Results

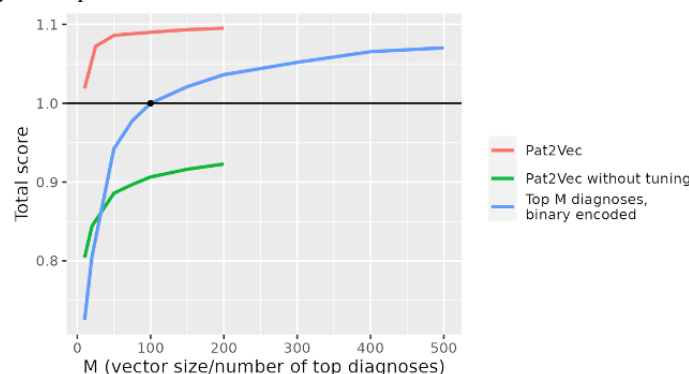
The Bayesian optimization search for the best hyperparameter configuration revealed that the default parameters are not sufficient and can be greatly improved upon ([Figure 2](#)). The

performance of the default parameter configuration did not exceed that of the top M diagnoses baseline model.

The most important hyperparameters (Supplementary Figure S1 in [Multimedia Appendix 1](#)) were (in order): the choice of DBOW over DM, the number of epochs (choosing 3), the negative sampling exponent (choosing approximately -2.3 , compared with the default [0.75]), and the learning rate alpha (choosing approximately 0.0014, compared with the default [0.025]).

When compared with the top M diagnoses approach with M=100, the final set of parameters with a vector size of 100 resulted in a 9 percent point increase on our aggregated performance metric. All final models with a vector size of 10 or larger increased performance over this baseline model of the top 100 diagnoses. For smaller vector sizes, the gains in performance compared with the baseline models of equal size were larger ([Figure 2](#)). After a vector size of about 50, the performance of the vectorization increased by lesser amounts.

Figure 2. A comparison of the default vectorization model, the baseline model (the top M diagnoses), and the final model after hyperparameter tuning based on the total score of how well they did on prediction tasks.



Linear/Logistic Regression Versus Gradient-Boosted Trees

The ensemble-based machine learning with LightGBM Regressor/Classifier on the final vectorization model performed better than the linear and logistic regression counterparts on the vectorization data as well as the top M diagnoses data (Supplementary Figure S2 in [Multimedia Appendix 1](#)).

Additionally, we observed a bigger increase in performance by switching from top M diagnoses data to Pat2Vec-derived vectors on smaller vector sizes, which stresses that information is compressed well by the vectorization. Furthermore, up to a vector size of about 100, the vectorization data with linear/logistic regression or LightGBM outperformed even the LightGBM approach on the binary-encoded data, which indicates that nonlinear properties of the patient profiles were

encoded in the vector embeddings. In summary, using gradient-boosted trees or vector embeddings is always beneficial, and the combination of the 2 yields the best results.

Robustness Analysis

Diagnosis Dropout

As a sensitivity or robustness analysis of the vector embedding (and the baseline binary encoding), we calculated total scores on the reduced dropout data (with 10%, 25%, and 50% of diagnosis codes missing, respectively). We observed a steeper decrease for the binary-encoded top 100 diagnoses data, while the performance of the vectorization suffers mildly even with a 50% drop out of the diagnosis data (Supplementary Figure S3 in [Multimedia Appendix 1](#)).

Vectorization Training Data Sample Size

As an additional robustness analysis of the vector embedding with regard to necessary training data size, we calculated total scores on reduced vectorization training data, from 100% (the original 10 million patients' training data) to 0.1% of the original training data, or 10,000 patients. We observed a total score above 1 (thus, above the performance of the binary-encoded baseline model) for sample sizes as low as 0.5% of the original data, or 50,000 patients (Supplementary Figure S4 in [Multimedia Appendix 1](#)), while sample sizes of at least 1 million patients are needed to achieve total scores close to the total score on the original data.

Analysis of Patient Embedding

For visualization purposes, we projected the final vectorization model with a vector size of 100 into 2 dimensions using the UMAP algorithm. This way we were able to illustrate the high-dimensional vectorization and patterns within the patients' cohort ([Figures 3 and 4](#)).

We observed a triangular shape in the vector space of the embedded patient profiles, with multiple regions of higher density. The 3 corner areas are (1) young patients of both genders with a low number of cases and low prescription costs; (2) women with an average age below the average age of the cohort and with low prescription costs and a medium number of cases; and (3) elderly patients of both genders with a high number of cases and high prescription costs ([Figure 3](#)). The

HDBSCAN clustering identified 14 clusters but showed that many patients are not easily mapped to a cluster (50.67%, 42,024/82,937, of test data; [Figure 4](#)).

A closer inspection of the clusters revealed interesting patterns in the subcohorts ([Figure 4](#) and [Table 2](#); also see [Multimedia Appendix 2](#) for further details). The clusters 5, 13, and 14 all have a mean age of almost 70 years or older, but differ in the share of females, mean number of cases, rate of emergency cases, and drug spending costs. Among these clusters, cluster 13 is the oldest with distinctive ICD-10 diagnoses of F03 (dementia) and R32 (urinary incontinence), along with a large number of patients who do not appear in 2019's data, which indicates a high mortality within cluster 13. Clusters 5 and 6 have the most distinctive diagnosis codes in the H52 section (refractive errors/eyesight), but differ in their average age. Clusters 1 and 2 are almost exclusively female and of around the same mean age, but cluster 1 has a higher share of emergencies, and overexpressed ICD code Z34 (supervision of normal pregnancy) and section O09 (duration of pregnancy) point to pregnancy. Clusters 11 and 8 are the 2 youngest clusters, where cluster 11 is mostly characterized by routine examinations and vaccinations (Z00.1: routine child health examination; Z23.8 and Z27.8: immunizations), whereas cluster 8 is characterized by developmental disorders of speech and language (F80.9 and F80.0). Patients in cluster 12 have the most common acute ambulatory diseases (J06.9: acute upper respiratory infection; A09.9: gastroenteritis/colitis; and R51: headache). The remaining clusters show the other most prominent public health concerns in the German ambulatory health care system: cluster 3 (hay fever/asthma), cluster 4 (hypothyroidism), cluster 7 (depressive disorders), cluster 9 (pinched nerve/back pain/disc disorders), and cluster 10 (diabetes type 2).

Regarding the explainability or backward interpretation of our embedding, we analyzed how specific ICD-10 diagnosis codes map onto the patient vector dimensions. A heatmap of the correlations between a subset of 60 diagnosis codes and the 100D embedding showed that similar disease concepts were mapped to the same vector dimensions in a blockwise manner (Supplementary Figure S5 in [Multimedia Appendix 1](#)). It also showed that disease information was spread out over multiple dimensions instead of being mapped to only 1 dimension as in binary encoding.

Figure 3. UMAP embedding of Pat2Vec, colored by age/gender/number of cases in 2019/emergency treatment in 2019/last available year in claims data/drug prescription costs in 2019. f: female; m: male; UMAP: uniform manifold approximation and projection.

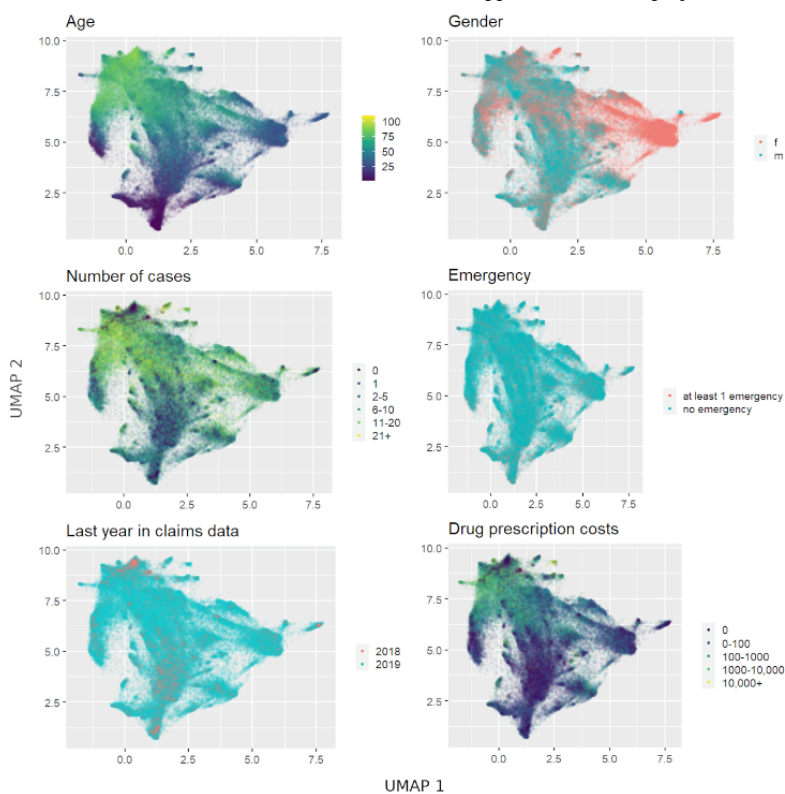


Figure 4. UMAP embedding of Pat2Vec, numbers 1-14 indicate clusters found by HDBSCAN (hierarchical density-based spatial clustering of applications with noise). UMAP: uniform manifold approximation and projection.

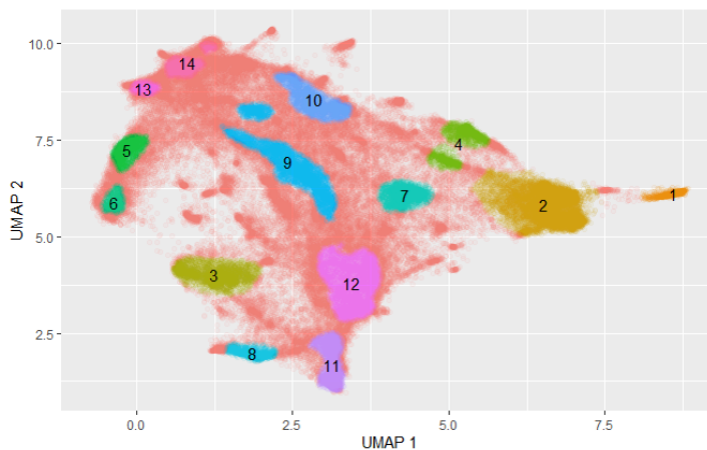


Table 2. Properties of clustered patients' cohorts.

Cluster	Percentage of cohort	Mean age (years)	Female, %	Mean number of cases	Emergency, %	Mean drug spending (€ ^a)	Distinctive ICD-10 ^b codes
11	3.8	4.1	50.4	4.8	35.2	69.26	Z00.1, Z23.8, Z27.8
8	1.5	9.4	35.9	5.7	27.1	198.01	F80.9, F80.0, Z00.1
6	1.1	21.7	49.0	5.3	21.8	62.77	H52.2, H52.0, H52.1
12	6.7	27.6	31.3	4.6	19.8	175.77	J06.9, A09.9, R51
1	1.7	32.0	99.9	8.4	28.4	230.47	Z34, N89.8, O09.3
3	4.0	33.3	38.1	7.1	19.1	323.30	J30.1, J45.9, J45.0
2	9.3	33.7	99.7	8.6	18.7	130.00	N89.8, Z30.9, Z12.9
7	2.6	44.5	57.1	9.9	19.0	431.01	F32.9, F32.1, F33.1
4	2.4	48.6	86.7	9.9	13.9	191.26	E03.9, E06.3, Z12.9
9	6.6	57.6	47.0	10.4	15.7	592.98	M54.1, M51.2, M54.5
10	3.7	59.3	37.3	8.4	11.5	480.11	I10.9, I10.90, E11.9
5	2.1	69.9	59.6	10.9	12.9	809.16	H52.2, H52.4, H52.0
14	2.6	74.4	37.4	11.9	16.0	1587.98	I10.9, I10.90, I25.1
13	1.3	80.7	62.9	8.2	26.6	1248.64	F03, R32, I10.9
None	50.7	50.2	51.0	9.4	17.9	908.89	N/A ^c
All	100.0	45.6	54.5	8.7	18.7	654.17	N/A

^a€=US \$1.08 (as of March 27, 2023).

^bICD-10: International Statistical Classification of Diseases and Related Health Problems, 10th Revision.

^cN/A: not applicable.

Prediction of Drug Spending Costs

Predicting prospective individual drug spending from diagnosis data is an especially hard task [66]. We predicted 2019's patient-level drug spending based on patients' diagnosis codes from 2018. We used and compared the binary-encoded top 100 diagnoses and our vectorization of dimension 100 (Pat2Vec). In addition, we extended the data by age and gender of patients. Table 3 shows the results using linear regression as well as gradient-boosted trees. We observed an overall high relative

increase in performance by using the vectorization over the baseline model, while in general the R^2 values were low. The linear regression shows diverging results between the top 100 and vectorization data with regard to absolute errors and squared errors (CPM and R^2). The gradient-boosted trees approach to regression performed similarly to the linear regression on the baseline model of binary-encoded top 100 diagnoses, while the combination of Pat2Vec and gradient-boosted trees performed best. Adding age and gender as additional variables led only to small increases in performance.

Table 3. R^2 , mean absolute error, and Cumming prediction measure of predicting drug spending costs using linear regression and LightGBM Regressor.

Measure	Linear regression			LightGBM Regressor		
	R^2 , %	Mean absolute error (€ ^a)	Cumming prediction measure, %	R^2 , %	Mean absolute error (€)	Cumming prediction measure, %
Age + gender	1.0	818.44	7.4	1.1	801.09	9.4
Top 100	2.0	760.55	14.0	2.1	755.76	14.5
Top 100 + age + gender	2.0	757.13	14.4	2.4	752.78	14.9
Pat2Vec	7.7	845.99	4.3	12.9	704.01	20.4
Pat2Vec + age + gender	7.7	845.98	4.3	13.7	690.70	21.9

^a€=US \$1.08 (as of March 27, 2023).

Discussion

Principal Findings

We found that the NLP-based vector embeddings of claims data led to large improvements on health care–related prediction tasks compared with standard approaches (represented by binary encoding). Hyperparameter tuning is necessary for these improvements. On health care prediction tasks, gradient-boosted tree algorithms outperform standard statistical methods (linear or logistic regression). Gradient-boosted trees benefit more from vectorization. Additionally, the performance of the vectorization is more robust against incomplete data, but at least 1 million patients are needed to train the vectorization model. Furthermore, our cohort analysis shows that most patients' diagnosis profiles lie on a spectrum of morbidity and cannot be easily mapped to distinct patient clusters. Overall, the results suggest we achieved the intended compression of the complete patient profiles while keeping the relevant amount of available information for prediction tasks.

Comparison With Previous Research

Embeddings of diagnosis codes have been studied extensively before [14,29,38-42]. Patient-level embeddings have been derived rarely [14,25,42]. To the best of our knowledge, there is no ICD-10–based patient vectorization model trained and optimized for application in generalized health care tasks.

Choi et al [39] trained ICD-9 code representations using another similar NLP approach, and at the same time they learned “visit representations” (vectors) based on a binary encoding of the diagnosis codes for individual visits. Using logistic regression and these representations of visits, they were able to predict future disease codes from 1 visit to the next and clinical risk groups [27]. In a similar way, Pham et al [41] trained diagnosis code representations and combined them into variable-size “admission representations” as input for a long short-term memory (LSTM) to predict individual health prognoses after a health care intervention.

Miotto et al [25] derived a patient-level embedding (Deep Patient) using autoencoders based on ICD-9 diagnosis codes in conjunction with medications, procedures, laboratory tests, clinical notes (free-text), and demographic variables. They used random forests and patient embeddings to predict future diseases, but they did not tune their embedding algorithm or prepare it for more general tasks.

Nguyen et al [42] found diagnosis code embeddings using Word2Vec. Subsequently, given an outcome, they trained a convolutional neural network to find predictive motifs for a classifier. They arrived at a patient-level embedding after the convolutional neural network step, but these embeddings are dependent on the classification task (they predicted unplanned readmissions in a hospital setting).

Almog et al [14] applied a similar approach (Crystal Bone) to the special problem of predicting bone fracture incidents. For the prediction of this specific task, they trained their vectorization models on data filtered for bone incidents. They described 2 approaches: gradient-boosted trees (using XGBoost [67]) on patients' vector embeddings as well as an LSTM [68]

neural network on the individual sequences of patients' diagnosis code embeddings. They observed better performance with the LSTM approach.

Li et al [29] derived an embedding for disease codes and a framework to predict diseases and even generalized outcomes (BEHRT). They did not set up a patient-level embedding with a fixed size, and their embedding framework needs to be retrained for new prediction tasks.

We were more interested in a general compression and embedding of patients themselves for general health care–related tasks (such as the prediction of different outcomes and an overall visualization) and not just the optimization of 1 prediction task only, thus we trained on the data of all patients, not filtered for specific diagnoses, and restricted ourselves to the analysis of the patients' vector embeddings. In addition, our embedding is based solely on the ICD-10 diagnosis data and does not need additional data sources that might not be readily available in a claims data setting. It would be helpful to look into how well other advanced machine learning algorithms such as LSTM or convolutional neural networks work on the ICD or patient vector embeddings for health care prediction tasks, but this is outside the scope of this paper.

Adkins [69] discussed the implications of a widespread adoption of machine learning on EHR data in clinical prediction contexts. While arguing that more complex machine learning models (such as the one presented in this work, combining vectorization and ensemble trees) on growing bodies of data will yield more precise predictions at the price of interpretability (as well as unforeseen ethical and legal issues), they pointed out the limitations of considering a limited amount of ICD codes, a problem that we could address to a large extent in our work. Interpreting the dimensions of the vectorizations and other steps to “explainable machine learning/artificial intelligence” are still ongoing (eg, building on the Shapley additive explanations values for tree methods [70,71]). Here, we employed a simple approach using correlations between vector embeddings and binary encoding to allow interpretation of vector dimensions with regard to specific ICD-10 codes.

Limitations and Strengths

It has been discussed that a fusion of EHR data (clinical/diagnosis data and laboratory quantitative measurements) and other data sources (eg, medical images and laboratory measurements) would lead to further advancements in health care prediction tasks [72,73], where the problems of these mixed data types need to be properly addressed. Unfortunately, the claims data of the presented analysis do not contain these additional data sources, and thus the current implementation cannot acknowledge this.

We set up access to a pretrained model of our vectorization with 10 dimensions so that other researchers in the field can evaluate our methods and use the model on their own health care data [74].

Future Research

The next step will be to use the provided vectorization for relevant tasks to improve health care. We will investigate

whether our approach will benefit tasks such as disease prediction with a long genesis time and prevention in cases of early detection, such as dementia and mild cognitive impairment. Furthermore, we will compare the benefits of data-driven vectorization with common EHR-based procedures such as the Elixhauser score [18] or clinical risk groups [27] in terms of describing patient cohorts or predicting health care outcomes. We think that patient clustering based on robust vectorization has the potential to identify patients who would benefit from early screening, which would lead to more personalized screening measures.

Conclusions

Health care–related prediction tasks that rely on large samples of data should make use of vectorization instead of binary encoding. Our fully pretrained and validated model can be used on new and possibly small data sets as well. Advanced machine learning techniques profit more from our vectorization. We enable more precise prediction models for decisions on future public health policies as well as more accurate health care services for individual patients.

Acknowledgments

This work is funded and contracted by the Associations of Statutory Health Insurance Physicians in the German Federal States.

Data Availability

The data sets that formed the training data during this study are not publicly available due to the regulations for sensitive health data in Article 9 of the General Data Protection Regulation (GDPR) of the European Union. Access can be given by official boards within the context of specific research projects, and the authors are available to discuss such possibilities. An embedding model that was made as part of this study is available online so that other researchers in the field can evaluate our procedures and apply the model to their own health care data [74].

Conflicts of Interest

This work and the Central Research Institute of Ambulatory Health Care in Germany (Zi) are funded and contracted by the Associations of Statutory Health Insurance Physicians in the German Federal States. It is its task to support and further develop the health care assurance mandate under German law.

Multimedia Appendix 1

Information on previous studies, top M diagnoses, hyperparameter importance, performance comparisons, and vector loadings. [PDF File (Adobe PDF File), 342 KB - [ai_v2i1e40755_app1.pdf](#)]

Multimedia Appendix 2

Extended Table 2 of main manuscript.

[PDF File (Adobe PDF File), 102 KB - [ai_v2i1e40755_app2.pdf](#)]

References

1. Häyrynen K, Saranto K, Nykänen P. Definition, structure, content, use and impacts of electronic health records: a review of the research literature. *Int J Med Inform* 2008 May;77(5):291-304. [doi: [10.1016/j.ijmedinf.2007.09.001](#)] [Medline: [17951106](#)]
2. Dean BB, Lam J, Natoli JL, Butler Q, Aguilar D, Nordyke RJ. Review: use of electronic medical records for health outcomes research: a literature review. *Med Care Res Rev* 2009 Dec;66(6):611-638. [doi: [10.1177/1077558709332440](#)] [Medline: [19279318](#)]
3. Jensen P, Jensen L, Brunak S. Mining electronic health records: towards better research applications and clinical care. *Nat Rev Genet* 2012 May 02;13(6):395-405. [doi: [10.1038/nrg3208](#)] [Medline: [22549152](#)]
4. Casey J, Schwartz B, Stewart W, Adler N. Using Electronic Health Records for Population Health Research: A Review of Methods and Applications. *Annu Rev Public Health* 2016;37:61-81 [FREE Full text] [doi: [10.1146/annurev-publhealth-032315-021353](#)] [Medline: [26667605](#)]
5. World Health Organization. *International Statistical Classification of Diseases and Related Health Problems, 10th Revision, Fifth Edition, 2016*. Geneva, Switzerland: World Health Organization; 2015.
6. Weiss J, Page D, Peissig PL, Natarajan S, McCarty C. Statistical Relational Learning to Predict Primary Myocardial Infarction from Electronic Health Records. *Proc Innov Appl Artif Intell Conf* 2012;2012:2341-2347 [FREE Full text] [Medline: [25360347](#)]
7. Cheng Y, Wang F, Zhang P, Hu J. Risk Prediction with Electronic Health Records: A Deep Learning Approach. 2016 Presented at: SIAM International Conference on Data Mining (SDM); May 5-7, 2016; Miami, FL p. 432-440. [doi: [10.1137/1.9781611974348.49](#)]

8. Choi E, Bahadori M, Kulas J, Schuetz A, Stewart W, Sun J. RETAIN: An interpretable predictive model for healthcare using REverse time AttentIoN mechanism. 2016 Presented at: 30th International Conference on Neural Information Processing Systems; December 5-10, 2016; Barcelona, Spain p. 3512-3520 URL: <https://dl.acm.org/doi/10.5555/3157382.3157490> [doi: [10.5555/3157382.3157490](https://doi.org/10.5555/3157382.3157490)]
9. Yu S, Chakraborty A, Liao K, Cai T, Ananthakrishnan A, Gainer V, et al. Surrogate-assisted feature extraction for high-throughput phenotyping. *J Am Med Inform Assoc* 2017 Apr 01;24(e1):e143-e149 [FREE Full text] [doi: [10.1093/jamia/ocw135](https://doi.org/10.1093/jamia/ocw135)] [Medline: [27632993](https://pubmed.ncbi.nlm.nih.gov/27632993/)]
10. Rubin KH, Möller S, Holmberg T, Bliddal M, Søndergaard J, Abrahamsen B. A New Fracture Risk Assessment Tool (FREM) Based on Public Health Registries. *J Bone Miner Res* 2018 Nov;33(11):1967-1979 [FREE Full text] [doi: [10.1002/jbmr.3528](https://doi.org/10.1002/jbmr.3528)] [Medline: [29924428](https://pubmed.ncbi.nlm.nih.gov/29924428/)]
11. Steele AJ, Denaxas SC, Shah AD, Hemingway H, Luscombe NM. Machine learning models in electronic health records can outperform conventional survival models for predicting patient mortality in coronary artery disease. *PLoS One* 2018;13(8):e0202344 [FREE Full text] [doi: [10.1371/journal.pone.0202344](https://doi.org/10.1371/journal.pone.0202344)] [Medline: [30169498](https://pubmed.ncbi.nlm.nih.gov/30169498/)]
12. Jorge A, Castro V, Barnado A, Gainer V, Hong C, Cai T, et al. Identifying lupus patients in electronic health records: Development and validation of machine learning algorithms and application of rule-based algorithms. *Semin Arthritis Rheum* 2019 Aug;49(1):84-90 [FREE Full text] [doi: [10.1016/j.semarthrit.2019.01.002](https://doi.org/10.1016/j.semarthrit.2019.01.002)] [Medline: [30665626](https://pubmed.ncbi.nlm.nih.gov/30665626/)]
13. Tomašev N, Glorot X, Rae JW, Zielinski M, Askham H, Saraiva A, et al. A clinically applicable approach to continuous prediction of future acute kidney injury. *Nature* 2019 Aug;572(7767):116-119 [FREE Full text] [doi: [10.1038/s41586-019-1390-1](https://doi.org/10.1038/s41586-019-1390-1)] [Medline: [31367026](https://pubmed.ncbi.nlm.nih.gov/31367026/)]
14. Almog YA, Rai A, Zhang P, Moulaison A, Powell R, Mishra A, et al. Deep Learning With Electronic Health Records for Short-Term Fracture Risk Identification: Crystal Bone Algorithm Development and Validation. *J Med Internet Res* 2020 Oct 16;22(10):e22550 [FREE Full text] [doi: [10.2196/22550](https://doi.org/10.2196/22550)] [Medline: [32956069](https://pubmed.ncbi.nlm.nih.gov/32956069/)]
15. Kogan E, Twyman K, Heap J, Milentijevic D, Lin J, Alberts M. Assessing stroke severity using electronic health record data: a machine learning approach. *BMC Med Inform Decis Mak* 2020 Jan 08;20(1):8 [FREE Full text] [doi: [10.1186/s12911-019-1010-x](https://doi.org/10.1186/s12911-019-1010-x)] [Medline: [31914991](https://pubmed.ncbi.nlm.nih.gov/31914991/)]
16. Martinez DA, Levin SR, Klein EY, Parikh CR, Menez S, Taylor RA, et al. Early Prediction of Acute Kidney Injury in the Emergency Department With Machine-Learning Methods Applied to Electronic Health Record Data. *Ann Emerg Med* 2020 Oct;76(4):501-514. [doi: [10.1016/j.annemergmed.2020.05.026](https://doi.org/10.1016/j.annemergmed.2020.05.026)] [Medline: [32713624](https://pubmed.ncbi.nlm.nih.gov/32713624/)]
17. Su C, Aseltine R, Doshi R, Chen K, Rogers SC, Wang F. Machine learning for suicide risk prediction in children and adolescents with electronic health records. *Transl Psychiatry* 2020 Nov 26;10(1):413 [FREE Full text] [doi: [10.1038/s41398-020-01100-0](https://doi.org/10.1038/s41398-020-01100-0)] [Medline: [33243979](https://pubmed.ncbi.nlm.nih.gov/33243979/)]
18. Elixhauser A, Steiner C, Harris DR, Coffey RM. Comorbidity measures for use with administrative data. *Med Care* 1998 Jan;36(1):8-27. [doi: [10.1097/00005650-199801000-00004](https://doi.org/10.1097/00005650-199801000-00004)] [Medline: [9431328](https://pubmed.ncbi.nlm.nih.gov/9431328/)]
19. Moore B, White S, Washington R, Coenen N, Elixhauser A. Identifying Increased Risk of Readmission and In-hospital Mortality Using Hospital Administrative Data: The AHRQ Elixhauser Comorbidity Index. *Med Care* 2017 Jul;55(7):698-705. [doi: [10.1097/MLR.0000000000000735](https://doi.org/10.1097/MLR.0000000000000735)] [Medline: [28498196](https://pubmed.ncbi.nlm.nih.gov/28498196/)]
20. Corey K, Kashyap S, Lorenzi E, Lagoo-Deenadayalan S, Heller K, Whalen K, et al. Development and validation of machine learning models to identify high-risk surgical patients using automatically curated electronic health record data (Pythia): A retrospective, single-site study. *PLoS Med* 2018 Nov;15(11):e1002701 [FREE Full text] [doi: [10.1371/journal.pmed.1002701](https://doi.org/10.1371/journal.pmed.1002701)] [Medline: [30481172](https://pubmed.ncbi.nlm.nih.gov/30481172/)]
21. Rahimian F, Salimi-Khorshidi G, Payberah A, Tran J, Ayala Solares R, Raimondi F, et al. Predicting the risk of emergency admission with machine learning: Development and validation using linked electronic health records. *PLoS Med* 2018 Nov;15(11):e1002695 [FREE Full text] [doi: [10.1371/journal.pmed.1002695](https://doi.org/10.1371/journal.pmed.1002695)] [Medline: [30458006](https://pubmed.ncbi.nlm.nih.gov/30458006/)]
22. Hippisley-Cox J, Coupland C. Predicting risk of emergency admission to hospital using primary care data: derivation and validation of QAdmissions score. *BMJ Open* 2013 Aug 19;3(8):e003482 [FREE Full text] [doi: [10.1136/bmjopen-2013-003482](https://doi.org/10.1136/bmjopen-2013-003482)] [Medline: [23959760](https://pubmed.ncbi.nlm.nih.gov/23959760/)]
23. Agresti A. *Categorical Data Analysis*, 3rd Edition. Hoboken, NJ: John Wiley & Sons; 2013.
24. Choi E, Bahadori M, Schuetz A, Stewart W, Sun J. Doctor AI: Predicting Clinical Events via Recurrent Neural Networks. *JMLR Workshop Conf Proc* 2016 Aug;56:301-318 [FREE Full text] [Medline: [28286600](https://pubmed.ncbi.nlm.nih.gov/28286600/)]
25. Miotto R, Li L, Kidd B, Dudley J. Deep Patient: An Unsupervised Representation to Predict the Future of Patients from the Electronic Health Records. *Sci Rep* 2016 May 17;6:26094 [FREE Full text] [doi: [10.1038/srep26094](https://doi.org/10.1038/srep26094)] [Medline: [27185194](https://pubmed.ncbi.nlm.nih.gov/27185194/)]
26. Yang S, Bian J, Sun Z, Wang L, Zhu H, Xiong H, et al. Early Detection of Disease Using Electronic Health Records and Fisher's Wishart Discriminant Analysis. *Procedia Computer Science* 2018;140:393-402. [doi: [10.1016/j.procs.2018.10.299](https://doi.org/10.1016/j.procs.2018.10.299)]
27. Hughes J, Averill R, Eisenhandler J, Goldfield N, Muldoon J, Neff J, et al. Clinical Risk Groups (CRGs): a classification system for risk-adjusted capitation-based payment and health care management. *Med Care* 2004 Jan;42(1):81-90. [doi: [10.1097/01.mlr.0000102367.93252.70](https://doi.org/10.1097/01.mlr.0000102367.93252.70)] [Medline: [14713742](https://pubmed.ncbi.nlm.nih.gov/14713742/)]

28. Ayala Solares JR, Diletta Raimondi FE, Zhu Y, Rahimian F, Canoy D, Tran J, et al. Deep learning for electronic health records: A comparative review of multiple deep neural architectures. *J Biomed Inform* 2020 Jan;101:103337 [FREE Full text] [doi: [10.1016/j.jbi.2019.103337](https://doi.org/10.1016/j.jbi.2019.103337)] [Medline: [31916973](https://pubmed.ncbi.nlm.nih.gov/31916973/)]
29. Li Y, Rao S, Solares J, Hassaine A, Ramakrishnan R, Canoy D, et al. BEHRT: Transformer for Electronic Health Records. *Sci Rep* 2020 Apr 28;10(1):7155 [FREE Full text] [doi: [10.1038/s41598-020-62922-y](https://doi.org/10.1038/s41598-020-62922-y)] [Medline: [32346050](https://pubmed.ncbi.nlm.nih.gov/32346050/)]
30. Goodfellow I, Bengio Y, Courville A. *Deep Learning*. Cambridge, MA: MIT Press; 2016.
31. ICD-10-GM Version 2020, Systematisches Verzeichnis, Internationale statistische Klassifikation der Krankheiten und verwandter Gesundheitsprobleme, 10. Revision, Stand: 20. September 2019. Köln, Germany: Deutsches Institut für Medizinische Dokumentation und Information (DIMDI) im Auftrag des Bundesministeriums für Gesundheit (BMG) unter Beteiligung der Arbeitsgruppe ICD des Kuratoriums für Fragen der Klassifikation im Gesundheitswesen (KKG); 2019. URL: <https://www.bfarm.de/EN/Code-systems/Classifications/ICD/ICD-10-GM/node.html> [accessed 2023-03-28]
32. Le Q, Mikolov T. Distributed representations of sentences and documents. 2014 Jun Presented at: 31st International Conference on Machine Learning; June 22-24, 2014; Beijing, China p. 1188-1196 URL: <https://dl.acm.org/doi/10.5555/3044805.3045025>
33. Mikolov T, Chen K, Corrado G, Dean J. Efficient Estimation of Word Representations in Vector Space. arxiv Preprint posted online on September 7, 2013 [FREE Full text] [doi: [10.48550/arXiv.1301.3781](https://doi.org/10.48550/arXiv.1301.3781)]
34. Mikolov T, Sutskever I, Chen K, Corrado G, Dean J. Distributed representations of words and phrases and their compositionality. 2013 Presented at: 27th Conference on Neural Information Processing Systems; December 5-10, 2013; Stateline, NV p. 3111-3119 URL: <https://arxiv.org/abs/1310.4546>
35. Caselles-Dupré H, Lesaint F, Royo-Letelier J. Word2vec applied to recommendation: Hyperparameters matter. New York, NY: ACM; 2018 Presented at: 12th ACM Conference on Recommender Systems; October 2-7, 2018; Vancouver, BC, Canada p. 352-356. [doi: [10.1145/3240323.3240377](https://doi.org/10.1145/3240323.3240377)]
36. Du J, Jia P, Dai Y, Tao C, Zhao Z, Zhi D. Gene2vec: distributed representation of genes based on co-expression. *BMC Genomics* 2019 Feb 04;20(Suppl 1):82 [FREE Full text] [doi: [10.1186/s12864-018-5370-x](https://doi.org/10.1186/s12864-018-5370-x)] [Medline: [30712510](https://pubmed.ncbi.nlm.nih.gov/30712510/)]
37. Chamberlain B, Rossi E, Shiebler D, Sedhain S, Bronstein M. Tuning Word2vec for large scale recommendation system. New York, NY: Association for Computing Machinery; 2020 Presented at: 14th ACM Conference on Recommender Systems; September 22-26, 2020; Virtual. [doi: [10.1145/3383313.3418486](https://doi.org/10.1145/3383313.3418486)]
38. Tran T, Nguyen TD, Phung D, Venkatesh S. Learning vector representation of medical objects via EMR-driven nonnegative restricted Boltzmann machines (eNRBM). *J Biomed Inform* 2015 Apr;54:96-105 [FREE Full text] [doi: [10.1016/j.jbi.2015.01.012](https://doi.org/10.1016/j.jbi.2015.01.012)] [Medline: [25661261](https://pubmed.ncbi.nlm.nih.gov/25661261/)]
39. Choi E, Bahadori M, Searles E, Coffey C, Thompson M, Bost J, et al. Multi-layer representation learning for medical concepts. New York, NY: Association for Computing Machinery; 2016 Presented at: 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining; August 13-17, 2016; San Francisco, CA p. 1495-1504. [doi: [10.1145/2939672.2939823](https://doi.org/10.1145/2939672.2939823)]
40. Choi Y, Chiu CYI, Sontag D. Learning Low-Dimensional Representations of Medical Concepts. *AMIA Jt Summits Transl Sci Proc* 2016;2016:41-50 [FREE Full text] [Medline: [27570647](https://pubmed.ncbi.nlm.nih.gov/27570647/)]
41. Pham T, Tran T, Phung D, Venkatesh S. DeepCare: a deep dynamic memory model for predictive medicine. In: Bailey J, Khan L, Washio T, Dobbie G, Huang JZ, Wang R, editors. *Advances in Knowledge Discovery and Data Mining : 20th Pacific-Asia Conference, PAKDD 2016 Auckland, New Zealand, April 19–22, 2016 Proceedings, Part II*. Cham, Switzerland: Springer; 2016:30-41.
42. Nguyen P, Tran T, Wickramasinghe N, Venkatesh S. A Convolutional Net for Medical Records. *IEEE J Biomed Health Inform* 2017 Jan;21(1):22-30. [doi: [10.1109/JBHI.2016.2633963](https://doi.org/10.1109/JBHI.2016.2633963)] [Medline: [27913366](https://pubmed.ncbi.nlm.nih.gov/27913366/)]
43. DESTATIS. Mitglieder und mitversicherte Familienangehörige der gesetzlichen Krankenversicherung am 1.7. eines Jahres (Anzahl). Gesundheitsberichterstattung des Bundes. 2022. URL: <https://www.gbe-bund.de/gbe/> [accessed 2023-03-28]
44. Frahm N, Peters M, Bätzing J, Ellenberger D, Akmatov MK, Haas J, et al. Treatment patterns in pediatric patients with multiple sclerosis in Germany—a nationwide claim-based analysis. *Ther Adv Neurol Disord* 2021;14:17562864211048336 [FREE Full text] [doi: [10.1177/17562864211048336](https://doi.org/10.1177/17562864211048336)] [Medline: [34646362](https://pubmed.ncbi.nlm.nih.gov/34646362/)]
45. Tikkanen R, Osborn R, Mossialos E, Djordjevic A, Wharton G. *International profiles of health care systems*. The Commonwealth Fund. London, UK: The Commonwealth Fund; 2020. URL: <https://www.commonwealthfund.org/international-health-policy-center/system-profiles> [accessed 2023-03-31]
46. Blümel M, Spranger A, Achstetter K, Maresso A, Busse R. Germany: Health System Review. *Health Syst Transit* 2020 Dec;22(6):1-272 [FREE Full text] [Medline: [34232120](https://pubmed.ncbi.nlm.nih.gov/34232120/)]
47. Young T, Hazarika D, Poria S, Cambria E. Recent Trends in Deep Learning Based Natural Language Processing [Review Article]. *IEEE Comput. Intell. Mag* 2018 Aug;13(3):55-75. [doi: [10.1109/mci.2018.2840738](https://doi.org/10.1109/mci.2018.2840738)]
48. Minaee S, Kalchbrenner N, Cambria E, Nikzad N, Chenaghlu M, Gao J. Deep Learning--based Text Classification: A Comprehensive Review. *ACM Comput. Surv* 2021 Apr 17;54(3):1-40. [doi: [10.1145/3439726](https://doi.org/10.1145/3439726)]
49. Bergstra J, Bardenet R, Bengio Y, Kégl B. Algorithms for hyper-parameter optimization. Red Hook, NY: Curran Associates; 2011 Presented at: 24th International Conference on Neural Information Processing Systems; December 12-17, 2011;

- Granada, Spain p. 2546-2554 URL: <https://proceedings.neurips.cc/paper/2011/file/86e8f7ab32cfd12577bc2619bc635690-Paper.pdf>
50. Ke G, Meng Q, Finley T, Wang T, Chen W, Ma W, et al. LightGBM: A highly efficient gradient boosting decision tree. Red Hook, NY: Curran Associates; 2017 Presented at: 31st International Conference on Neural Information Processing Systems; December 4-9, 2017; Long Beach, CA p. 3149-3157 URL: <https://proceedings.neurips.cc/paper/2017/file/6449f44a102fde848669bdd9eb6b76fa-Paper.pdf>
 51. Friedman J. Greedy function approximation: A gradient boosting machine. *The Annals of Statistics* 2001;29(5):1189-1232. [doi: [10.1214/aos/1013203451](https://doi.org/10.1214/aos/1013203451)]
 52. Hastie T, Tibshirani R, Friedman J. Boosting and additive trees. In: *The elements of statistical learning: Data mining, inference, and prediction*. New York, NY: Springer; 2009:337-387.
 53. Al Daoud E. Comparison between XGBoost, LightGBM and CatBoost using a home credit dataset. *International Journal of Information, Control and Computer Sciences* 2019;13(1):6-10. [doi: [10.5281/zenodo.3607805](https://doi.org/10.5281/zenodo.3607805)]
 54. Bentéjac C, Csörgő A, Martínez-Muñoz G. A comparative analysis of gradient boosting algorithms. *Artif Intell Rev* 2020 Aug 24;54(3):1937-1967. [doi: [10.1007/s10462-020-09896-5](https://doi.org/10.1007/s10462-020-09896-5)]
 55. Cumming R, Knutson D, Cameron B, Derrick B. A comparative analysis of claims-based methods of health risk assessment for commercial populations: Final report to the Society of Actuaries. Society of Actuaries. 2002. URL: <https://www.soa.org/globalassets/assets/Files/Research/Projects/risk-assessmentc.pdf> [accessed 2023-03-31]
 56. Friedman J, Hastie T, Tibshirani R. Additive logistic regression: a statistical view of boosting (with discussion and a rejoinder by the authors). *The Annals of Statistics* 2000;28(2):337-407. [doi: [10.1214/aos/1016218223](https://doi.org/10.1214/aos/1016218223)]
 57. McInnes L, Healy J, Melville J. UMAP: Uniform manifold approximation and projection for dimension reduction. *Journal of Open Source Software* 2018;29(3):861. [doi: [10.21105/joss.00861](https://doi.org/10.21105/joss.00861)]
 58. Campello R, Moulavi D, Sander J. Density-based clustering based on hierarchical density estimates. Berlin/Heidelberg, Germany: Springer; 2013 Presented at: PAKDD 2013: Advances in Knowledge Discovery and Data Mining; April 14-17, 2013; Gold Coast, QLD, Australia p. 160-172. [doi: [10.1007/978-3-642-37456-2_14](https://doi.org/10.1007/978-3-642-37456-2_14)]
 59. Van RG, Drake F. Python 3 Reference Manual. Scotts Valley, CA: CreateSpace; 2009.
 60. R: A Language and Environment for Statistical Computing. The R Foundation. Vienna, Austria: R Foundation for Statistical Computing; 2020. URL: <https://www.R-project.org/> [accessed 2023-03-31]
 61. Řehůřek R, Sojka P. Software Framework for Topic Modelling with Large Corpora. 2010 Presented at: LREC 2010 Workshop New Challenges for NLP Frameworks; May 22, 2010; Valletta, Malta p. 46-50.
 62. Akiba T, Sano S, Yanase T, Ohta T, Koyama M. Optuna: A next-generation hyperparameter optimization framework. New York, NY: Association for Computing Machinery; 2019 Presented at: 25th ACM SIGKDD international conference on knowledge discovery and data mining; August 4-8, 2019; Anchorage, AK p. 2623-2631. [doi: [10.1145/3292500.3330701](https://doi.org/10.1145/3292500.3330701)]
 63. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, et al. Scikit-learn: Machine Learning in Python. *The Journal of Machine Learning Research* 2011 Nov 1;12:2825-2830 [FREE Full text] [doi: [10.5555/1953048.2078195](https://doi.org/10.5555/1953048.2078195)]
 64. Wickham H. ggplot2: Elegant Graphics for Data Analysis. New York, NY: Springer; 2016.
 65. Starker A, Buttman-Schweiger N, Krause L, Barnes B, Kraywinkel K, Holmberg C. [Cancer screening in Germany: availability and participation]. *Bundesgesundheitsblatt Gesundheitsforschung Gesundheitsschutz* 2018 Dec;61(12):1491-1499. [doi: [10.1007/s00103-018-2842-8](https://doi.org/10.1007/s00103-018-2842-8)] [Medline: [30406892](https://pubmed.ncbi.nlm.nih.gov/30406892/)]
 66. Zhao Y, Ash AS, Ellis RP, Ayanian JZ, Pope GC, Bowen B, et al. Predicting pharmacy costs and other medical costs using diagnoses and drug claims. *Med Care* 2005 Jan;43(1):34-43. [Medline: [15626932](https://pubmed.ncbi.nlm.nih.gov/15626932/)]
 67. Chen T, Guestrin C. XGBoost: A scalable tree boosting system. 2016 Presented at: 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining; August 13-17, 2016; San Francisco, CA p. 785-794 URL: <https://arxiv.org/abs/1603.02754> [doi: [10.1145/2939672.2939785](https://doi.org/10.1145/2939672.2939785)]
 68. Hochreiter S, Schmidhuber J. Long short-term memory. *Neural Comput* 1997 Nov 15;9(8):1735-1780. [doi: [10.1162/neco.1997.9.8.1735](https://doi.org/10.1162/neco.1997.9.8.1735)] [Medline: [9377276](https://pubmed.ncbi.nlm.nih.gov/9377276/)]
 69. Adkins DE. Machine Learning and Electronic Health Records: A Paradigm Shift. *Am J Psychiatry* 2017 Feb 01;174(2):93-94 [FREE Full text] [doi: [10.1176/appi.ajp.2016.16101169](https://doi.org/10.1176/appi.ajp.2016.16101169)] [Medline: [28142275](https://pubmed.ncbi.nlm.nih.gov/28142275/)]
 70. Lundberg S, Lee S. A unified approach to interpreting model predictions. Red Hook, NY: Curran Associates; 2017 Presented at: 31th Conference on Neural Information Processing Systems; December 4-9, 2017; Long Beach, CA URL: <https://proceedings.neurips.cc/paper/2017/file/8a20a8621978632d76c43dfd28b67767-Paper.pdf>
 71. Lundberg SM, Erion G, Chen H, DeGrave A, Prutkin JM, Nair B, et al. From Local Explanations to Global Understanding with Explainable AI for Trees. *Nat Mach Intell* 2020 Jan;2(1):56-67 [FREE Full text] [doi: [10.1038/s42256-019-0138-9](https://doi.org/10.1038/s42256-019-0138-9)] [Medline: [32607472](https://pubmed.ncbi.nlm.nih.gov/32607472/)]
 72. Huang S, Pareek A, Seyyedi S, Banerjee I, Lungren M. Fusion of medical imaging and electronic health records using deep learning: a systematic review and implementation guidelines. *NPJ Digit Med* 2020;3:136 [FREE Full text] [doi: [10.1038/s41746-020-00341-z](https://doi.org/10.1038/s41746-020-00341-z)] [Medline: [33083571](https://pubmed.ncbi.nlm.nih.gov/33083571/)]
 73. Tayefi M, Ngo P, Chomutare T, Dalianis H, Salvi E, Budrionis A, et al. Challenges and opportunities beyond structured data in analysis of electronic health records. *WIREs Comp Stat* 2021 Feb 14;13(6):e1549. [doi: [10.1002/wics.1549](https://doi.org/10.1002/wics.1549)]
 74. Steiger E, Kroll LE. Pat2Vec. Hugging Face. URL: <https://huggingface.co/zidatasciencelab/Pat2Vec> [accessed 2023-03-27]

Abbreviations

CPM: Cumming predictive measure

DBOW: distributed bag of words

DM: distributed memory

EHR: electronic health record

GDPR: General Data Protection Regulation

HDBSCAN: hierarchical density-based spatial clustering of applications with noise

ICD-10: International Statistical Classification of Diseases and Related Health Problems, 10th Revision

ICD-10-GM: International Statistical Classification of Diseases and Related Health Problems, 10th revision, German Modification

LightGBM: light gradient-boosted machine

LSTM: long short-term memory

NLP: natural language processing

SHI: statutory health insurance

UMAP: uniform manifold approximation and projection

Edited by K El Emam, B Malin; submitted 04.07.22; peer-reviewed by S Sarejloo, MS Aslam, SD Boie, W Zhang; comments to author 15.11.22; revised version received 09.12.22; accepted 18.03.23; published 21.04.23.

Please cite as:

Steiger E, Kroll LE

Patient Embeddings From Diagnosis Codes for Health Care Prediction Tasks: Pat2Vec Machine Learning Framework

JMIR AI 2023;2:e40755

URL: <https://ai.jmir.org/2023/1/e40755>

doi: [10.2196/40755](https://doi.org/10.2196/40755)

PMID:

©Edgar Steiger, Lars Eric Kroll. Originally published in JMIR AI (<https://ai.jmir.org>), 21.04.2023. This is an open-access article distributed under the terms of the Creative Commons Attribution License (<https://creativecommons.org/licenses/by/4.0/>), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR AI, is properly cited. The complete bibliographic information, a link to the original publication on <https://www.ai.jmir.org/>, as well as this copyright and license information must be included.

Original Paper

Prediction of Chronic Stress and Protective Factors in Adults: Development of an Interpretable Prediction Model Based on XGBoost and SHAP Using National Cross-sectional DEGS1 Data

Arezoo Bozorgmehr¹, MSc; Birgitta Weltermann¹, MD, MPH

Institute of General Practice and Family Medicine, University Hospital Bonn, University of Bonn, Bonn, Germany

Corresponding Author:

Arezoo Bozorgmehr, MSc

Institute of General Practice and Family Medicine

University Hospital Bonn

University of Bonn

Venusberg-Campus 1

Bonn, 53127

Germany

Phone: 49 228 287 11160

Email: arezoo.bozorgmehr@ukbonn.de

Abstract

Background: Chronic stress is highly prevalent in the German population. It has known adverse effects on mental health, such as burnout and depression. Known long-term effects of chronic stress are cardiovascular disease, diabetes, and cancer.

Objective: This study aims to derive an interpretable multiclass machine learning model for predicting chronic stress levels and factors protecting against chronic stress based on representative nationwide data from the German Health Interview and Examination Survey for Adults, which is part of the national health monitoring program.

Methods: A data set from the German Health Interview and Examination Survey for Adults study including demographic, clinical, and laboratory data from 5801 participants was analyzed. A multiclass eXtreme Gradient Boosting (XGBoost) model was constructed to classify participants into 3 categories including low, middle, and high chronic stress levels. The model's performance was evaluated using the area under the receiver operating characteristic curve, precision, recall, specificity, and the F_1 -score. Additionally, SHapley Additive exPlanations was used to interpret the prediction XGBoost model and to identify factors protecting against chronic stress.

Results: The multiclass XGBoost model exhibited the macroaverage scores, with an area under the receiver operating characteristic curve of 81%, precision of 63%, recall of 52%, specificity of 78%, and F_1 -score of 54%. The most important features for low-level chronic stress were male gender, very good general health, high satisfaction with living space, and strong social support.

Conclusions: This study presents a multiclass interpretable prediction model for chronic stress in adults in Germany. The explainable artificial intelligence technique SHapley Additive exPlanations identified relevant protective factors for chronic stress, which need to be considered when developing interventions to reduce chronic stress.

(JMIR AI 2023;2:e41868) doi:[10.2196/41868](https://doi.org/10.2196/41868)

KEYWORDS

artificial intelligence; machine learning; prognostic; model; chronic stress; resilience factors; interpretable model; explainability; stress; disease; diabetes; cancer; dataset; clinical; data; gender; social support; support; intervention; SHAP

Introduction

Chronic stress has many negative effects, primarily on mental health, for example burnout and depression [1]. Long-term chronic stress is associated with various illnesses including cardiovascular disease, diabetes, cancer, and asthma [2-5]. High

chronic stress is prevalent with multiple mental health problems in the German population, and this value has increased to 61.1% [6]. However, the vast majority of the population does not develop high chronic stress. While most research has focused on the development of pathology and risk factors, it is paramount to better understand protective factors that prevent chronic stress. In our prior study [7] with 764 participants including general

practitioners (GPs) and practice assistants (PrAs) from 136 German general practices, we analyzed the level of strain due to stress stratified for personal, practice, and regional characteristics. We showed that GPs and PrAs, who individually applied more than 5 measures regularly to compensate for stress, had markedly lower stress levels as measured by the Screening Scale of the Trier Inventory for the Assessment of Chronic Stress (TICS-SSCS) instrument [8].

The psychological construct of resilience, developed over the last decades, addresses this perspective. The American Psychological Association (in 2014) defines resilience as “the process of adapting well in the face of adversity, trauma, tragedy, threats or even significant sources of stress” [9]. Resilience in the context of chronic stress has been characterized by the ability to “bounce back from negative emotional experiences and by flexible adaptation to the changing demands of stressful experiences” [10]. It involves the ability to maintain healthy functioning in different domains of life, such as work and family. Holz et al [11] provided an overview of the current literature investigating the neural mechanisms of resilience focusing on social background. They discussed possible prevention and early intervention approaches targeting the individual and the social environment to lower the risk of psychiatric disorders and to foster resilience [11]. Schetter et al [12] reviewed the traditions of research and definitions of resilience to chronic stress in adults and gained an understanding of resilience in general. They developed a taxonomy of resilience resources to guide future research [12]. Other studies focused on neurobiological cascades involving, for example, enkephalins and associated opioid receptors, μ -opioid peptide receptor, and δ -opioid peptide receptor, to better understand the biological mechanisms of natural adaptation. Prospectively, this bares the potential for effective preventive or therapeutic strategies [13].

To better understand the chronic stress in epidemiological studies, machine learning (ML) offers new approaches to evaluate and model complex relationships in data [14,15]. ML strategies are based on algorithms, which describe the relationships between variables. Two areas in medicine that benefit from ML techniques are diagnosis and outcome prediction [16,17]. Focusing on chronic stress prediction, our prior study [18] compared 4 supervised ML classifiers and 1 standard approach based on data of 550 PrAs from 136 German general practices. We showed that all 4 ML approaches, especially random forest, provided more accurate models for predicting chronic stress than standard regression analysis [18].

Aiming at an interpretable multiclass ML model for predicting chronic stress, we developed an eXtreme Gradient Boosting (XGBoost) model based on nationally representative German Health Interview and Examination Survey for Adults (DEGS1) data. The unified framework SHAP (SHapley Additive exPlanations) is used to interpret the prediction model and to identify factors protecting against chronic stress.

Methods

Overview

This study used nationally representative data from the DEGS1 study, which is a part of the health monitoring program of the Robert Koch Institute, Berlin, Germany. It was conducted from 2008 to 2011 by means of interviews, examinations, and tests among the German population aged 18-79 years (n=8151). The DEGS1 data set, which is available for public use on request, included measurements for chronic stress among 5801 respondents aged 18 to 64 years [6,19].

Primary Outcome

Chronic stress was assessed using the 12-item German short version of TICS-SSCS (n=5850) [6]. It was developed by Schultz et al [8] based on the systemic-requirement-resource model of health [8,20]. The 12-item scale addresses 5 stress areas: chronic worrying, work overload, social overload, excessive demands of work, and lack of social recognition. Its internal consistency showed a Cronbach α of .91 and a good to very good reliability with values ranging from .84 to .91 (mean α =.87) [8]. All 12 questionnaire items use a 5-point Likert scale answer format (0=“never” to 4=“very often”) to measure chronic stress in the past 3 months [21,22]. A sum score (scale 0-48) was calculated for each participant, which is categorized in 3 classes based on a reference population with the TICS-SSCS: 1-11 (\leq median)=low stress, 12-22=middle stress, and >22 =high stress (\geq 90th percentile). This multiclass outcome is the recommended DEGS1 approach [6].

Predictors

In addition, the DEGS1 data set included variables on sociodemographic characteristics, chronic diseases (eg, coronary heart disease, stroke, diabetes mellitus, depression, and anxiety disorder), living conditions, health-related behavior, preventive measures, and general health. Based on a literature review and using the Powershap feature selection method, 34 features were included in this analysis. Table 1 depicts descriptive information about the variables used.

Table 1. Demographic, clinical, and workplace characteristics of the German Health Interview and Examination Survey for Adults study participants (N=5801).

Demographic characteristics	Values
Continuous variables, mean (SD; range)	
Age (years)	42 (13.11; 18-64)
Number of persons in the household	3 (1.34; 1-11)
Sleep hours per night in the past 4 weeks	7 (1.19; 2-12)
Number of hospital nights in the past 12 months	1 (5.30; 0-150)
Number of sick days in the past 12 months	13 (38.01; 0-365)
Categorical variables	
Gender (female), n (%)	3081 (49.6)
Marital status, n (%)	
Married living with partner or separately from partner	3697 (59.5)
Single	1957 (31.5)
Divorced	376 (6.1)
Widowed	136 (2.2)
Provides care to someone in need or seriously ill, n (%)	379 (6.1)
Renting or living in own apartment/house, n (%)	
Rented apartment or house	2689 (43.3)
Own apartment or house	3268 (52.6)
Satisfaction with living space, n (%)	
Very satisfied or satisfied	5269 (84.8)
Neither satisfied nor dissatisfied	608 (9.8)
Dissatisfied or very dissatisfied	295 (4.8)
Residential area satisfaction, n (%)	
Very satisfied or satisfied	5091 (81.9)
Neither satisfied nor dissatisfied	727 (11.7)
Dissatisfied or very dissatisfied	320 (5.2)
General state of health, n (%)	
Very good or good	4942 (79.5)
Average	1134 (18.3)
Poor or very poor	116 (1.8)
Intake of sleeping pills in the past 4 weeks, n (%)	
Never	5919 (95.3)
Less than 1 time	100 (1.6)
1 time or 2 times	73 (1.2)
3 times or more	86 (1.4)
Social support, n (%)	
Low support	653 (10.5)
Average support	3082 (49.6)
Strong support	2451 (39.5)
Health behavior consultation in the past 12 months, n (%)	
Has general practitioner	5497 (88.5)
Visited to general practitioner in the past 12 months	4870 (78.4)

Demographic characteristics	Values
Visited to neurologist in the past 12 months	463 (7.5)
Frequency of alcohol consumption, n (%)	
Never	744 (12.0)
1 time per month or less	1186 (19.1)
2-4 times per month	1998 (32.2)
2-3 times per week	1453 (23.4)
4 times per week or more	811 (13.1)
Tobacco use, n (%)	
Yes, daily	1701 (27.4)
Yes, occasionally	433 (7)
Not anymore	1664 (26.8)
Never smoked	2400 (38.7)
Comorbidities, n (%)	
Has hypertension	1625 (26.2)
Has diabetes	271 (4.4)
Has migraine	712 (11.5)
Has depression	682 (11)
Has anxiety disorder	327 (5.3)
Has burnout syndrome	292 (4.7)
Has one or more long-term chronic diseases	1418 (22.8)
Prevention programs or sport activities, n (%)	
Participated in prevention program in the past 12 months	988 (15.9)
Participated in relaxation or stress management program	188 (3)
Participated in gymnastics, fitness, or balance sports program	832 (13.4)
Participated in alcohol cessation program	7 (0.1)
Participated in smoking cessation program	17 (0.3)
Participated in weight reduction or a healthy diet program	167 (2.7)
Sports activities per week (in the past 3 months), n (%)	
No sports activity	1954 (31.5)
Up to 2 hours per week	2584 (41.6)
Regularly, 2-4 hours per week	990 (15.9)
Regularly, more than 4 hours per week	645 (10.4)

Data Preprocessing

Data Normalization

The DEGS1 study features include both discrete and continuous values. When these features are combined, the range of the values differs. Therefore, the training data set was normalized using the min-max normalization method. This normalization technique accurately preserves all relationships in the data, thereby avoiding the introduction of bias [23].

Handling of Missing Data

For single features, missing values were low (<2%), yielding an overall missing rate of 13.91% in our data set. We used the

K-Nearest Neighbors (KNN) approach to impute the missing variables. This method identifies the KNNs on the Euclidean distance. Missing values were replaced using a majority vote for discrete variables and weighted means for continuous features. All features are imputed simultaneously without the need to treat features individually [24].

Addressing the Imbalanced Data Set

For chronic stress, the distribution of classes was unequal (class 0: 52%, class 1: 38%, and class 2: 11%). This imbalanced multiclass classification was addressed using the Synthetic Minority Oversampling TEchnique to increase the frequency of near-miss data points within the training data set. This oversampling method randomly generated new instances of

minority class to balance the number of classes without any additional information to the model [25].

Feature Selection

We used Powershap as a wrapper-based Shaply feature selection method. This technique is based on the core assumption that an informative feature will have a larger impact on the prediction compared to a known random feature [26].

Machine Learning Approach: XGBoost

Overview

To predict chronic stress levels and detect factors protecting against chronic stress, we applied the decision tree-based ensemble ML technique, XGBoost [27,28]. XGBoost is a scalable and accurate implementation gradient boosting machine developed by the Distributed Machine Learning Community in the form of open-source libraries. It combines a recursive gradient boosting method called Newton boosting. Based on a decision tree model, it efficiently provides accurate predictions because each tree is boosted recursively and in parallel.

The ML technique generally aims to identify a relationship between the input $X = \{x_1, x_2, \dots, x_n\}$ and the output Y . For a given data set with n samples and m features, K additive functions are used in the XGBoost model to predict the output through the following estimation (equation 1) [27]:

Table 2. Main hyperparameters for the Extreme Gradient Boosting model.

Hyperparameter	Value
learning rate	0.3
Estimators, n	1000
max_depth	5
Subsample	0.8
min_child_weight	3
L2 regularization term (Lambda)	2
colsample-bytree	0.7
Objective	multi:softmax

K-Fold Cross-Validation

After preprocessing, the 34 features were fed into ML classifiers to train the model for classification. The data set was split into a “training” and a “validation” data set. We used the repeated K-fold cross-validation approach, repeating the mean performance across all folds and all repeats to reduce the bias in the model's estimated performance with $K=10$. $K=10$ was chosen as the optimal number of folds, which optimizes the time to complete the test while minimizing the bias and variance associated with the validation process.

Model Performance Evaluation

To evaluate the method proposed in this study, we used the following most promising multiclass evaluation metrics: the area under the receiver operating characteristic curve (AUC), precision, recall, and F_1 -score. Multiclass classification works on data sets in which all classes are mutually exclusive. In a



where $f_k = \{f(x) = \omega_q\}$ ($q: R^m \rightarrow T, \omega \in R^T$) is the regression tree's space, and q denotes the independent structure of each tree with T leaves. Each f_k corresponds to an independent tree structure q and leaf weights ω . The following regularized objective is minimized to learn the set of functions (equation 2).



where $\Omega(f) = \gamma T + \frac{1}{2} \lambda \|\omega\|^2$, I represents the model loss function, and Ω denotes the regularized term.

Hyperparameter Tuning

In this study, a grid-search approach from scikit-learn class “GridSearchCV” was applied toward the optimal tuning of XGBoost hyperparameters. The number of estimators was set to 1000 to represent the maximum number of trees created during the training phase. The Softmax function is used to convert logits of the XGBoost classifier into a probability distribution. Each element of the output lies in the interval (0,1) and the output elements sum up to 1. Table 2 summarizes the hyperparameters' values used to the XGBoost model (see Multimedia Appendix 1).

multiclass classifier, the evaluation measures of individual classes are averaged out to determine the performance on overall system across the data. We applied the macroaverage approach [29].

The receiver operating characteristic (ROC) curve was used to evaluate the performance of the classifier. For different classification thresholds, the macro true-positive rate (equation 3) is plotted against the macro false-positive rate (equation 4). The AUC indicates the classifier's ability to distinguish between classes. The value of the AUC is in the range (0,1), in which 1 is for a perfect classifier. In this study, the ROC curve is plotted for each class broken down into a series of binary problems using the One-vs-Rest approach. The macroaverage is computed by summing the individual values for true positive, true negative, false positive, and false negative. Then, macroaverage scores of true positive instances (precision; equation 5), true positive rate (recall; equation 6), true negative rate (specificity; equation 7), and the harmonic mean of the precision and recall computed

on each class (F_1 -score; equation 8) were computed. Mathematically, they are defined as follows:



We used Python 3.7 (Python Software Foundation) to implement our ML framework. In addition, several libraries from the python data science ecosystem were used to execute the experiments and the integrated development environment PyCharm. To implement the Powershap feature selection method, we used the Powershap Python library. The scikit-learn package (version 1.0.2) was used to train and evaluate the ML classifier. SHAP tool (version 0.40.0) was used to assess the explainability the model; that is, to identify factors protecting against chronic stress.

In addition to the performance evaluation, this study maximizes the interpretability of the underlying models. It focuses particularly on the explainability of the model, which can serve as an indispensable tool in the era of precision medicine.

Model Interpretation: SHAP

Per our understanding, the interpretation of the prediction models is as crucial as the prediction accuracy because it extracts information that significantly affects outcomes and identifies the factors protecting against chronic stress from subjects with lower chronic stress. However, the ensemble learning method XGBoost represents a black-box model. To overcome this problem, Lundberg [30,31] proposes the SHAP approach for interpreting predictions of complex models created by different techniques; for example, NGBoost, CatBoost, XGBoost, LightGBM, and scikit-learn tree models. SHAP was initially developed by Shapley in 1953 and is based on the game theory [32]. It explains the prediction of a specific input (\mathbf{X}) by calculating the impact of each feature on the prediction. The estimated Shapley values are calculated as follows (equation 9):



where \hat{f}_x is the prediction for x , but with a random number of feature values. TreeSHAP is used for gradient boosting models including XGBoost. It offers a rich visualization of each feature attribution and allows for partial dependence plots.

The TreeSHAP interaction values estimates as follows (equation 10):



where $i \neq j$, $\delta_{ij}(S) = f_x(S \cup \{i,j\}) - f_x(S \cup \{i\}) - f_x(S \cup \{j\}) + f_x(S)$, M is the number of features, and S denotes all feature subsets. SHAP values advance the understanding of tree models by including feature importance, feature dependence plots, local explanations, and summary plots [30].

Ethical Considerations

Ethics approval for the DEGS1 survey was obtained from the Charité – Universitätsmedizin Berlin Ethics Committee (EA2/047/08). All participants received written information and provided informed consent before the interview and examination. The analysis described here builds on a data set from the DEGS1 study, which was kindly provided by the Robert Koch Institute. This secondary analysis of anonymized data does not require a separate ethics vote.

Results

Characteristics of the DEGS1 Study Population

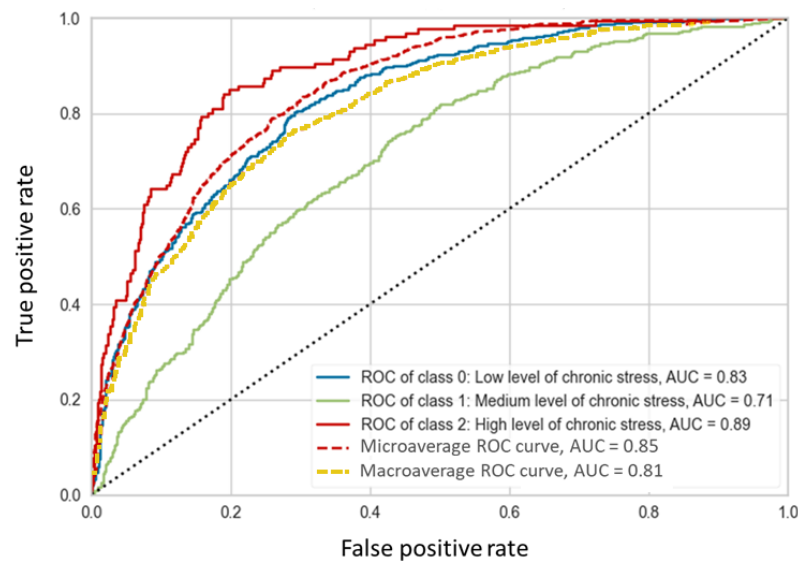
The mean age of the 5801 DEGS1 study participants was 44 years, with more than half of the population being female ($n=3080$, 53.1%). The mean stress level of the total population was 12.00 (95% CI 11.79-12.20): 11% ($n=625$) of the participants had “high chronic stress” (category 2), while 38% ($n=2188$) had “middle” (category 1), and 52% ($n=2988$) of them had “low chronic stress” (category 0). Most participants reported their general state of health as very good or good (79.3%, $n=4599$). Table 1 shows the weighted demographic, clinical, and laboratory characteristics of the participants.

Results of the Machine Learning Analysis

The evaluation metrics of the XGBoost model’s performance are presented in Table 3 differentiated by chronic stress classes. We see that the XGBoost model achieved the highest AUC score for class 2 with 0.89% and a good macroaverage AUC score of 81% for the overall model. The metrics for the 3 stress classes and the average results are reported in Table 3. The ROC curves for the multiclass chronic stress prediction of the XGBoost model are shown in Figure 1.

Table 3. Classification metrics: area under the receiver operating characteristic curve (AUC), precision, recall, specificity, and F_1 -score for XGBoost.

Measure	XGBoost			
	Class 0	Class 1	Class 2	Macroaverage
AUC	0.83	0.71	0.89	0.81
Precision	0.73	0.56	0.58	0.63
Recall	0.80	0.55	0.37	0.52
Specificity	0.90	0.38	0.26	0.78
F_1 -score	0.76	0.60	0.45	0.54

Figure 1. ROC curves for 3 classes using the XGBoost multiclass classifier. AUC: area under the receiver operating characteristic curve; ROC: receiver operating characteristic curve.

Explanation of the Behavior of Individual Features

The result of the SHAP analysis is displayed in Figure 2. In this plot, the impact of a feature on the respective classes (stress classes 0-2) is stacked to illustrate the feature importance. This means that the features with large absolute Shapley values are more important than those with lower values. The plot shows that class 0 (low level of chronic stress) hardly uses the features gender, general state of health, satisfaction with living space, and social support. Class 2 as the high level of chronic stress uses the features number of sick days in the past 12 months, social support, sleeping hours per night in the past 4 weeks, gender, and general state of health. Interestingly, classes 0 and 2 use many identical features.

While the SHAP feature plot provides an overview of the role of each variable irrespective of the direction of these effects, the SHAP summary plot provides such additional information for classes. The impact distribution of each feature on the model output for classes with low and high levels of chronic stress is shown in Figures 3 and 4. Each row in this plot represents a single feature in order of their mean absolute SHAP values. It can be a negative or positive value and represents the importance

of each feature. Each dot is a Shapley value for a particular feature and reflects its impact on a specific class for a given instance, and dots stack up to show density. It is color-coded in accordance with the magnitude to which the value contributes to the model impact (red=high and blue=low). The color is the actual feature value in the data set. For example, the red values for age as a continuous feature represent older people, while blue values represent younger people, and blue values for gender as a categorical feature (low value=1) represent males and red values (high value= 2) represent females. Overlapping points are jittered toward the y-axis, giving a sense of the distribution of the Shapley values per feature.

According to the SHAP summary plot result, gender is the most significant feature for class 0, and the number of sick days in the past 12 months has the highest impact on class 2. We note that the general state of health (shown in red) with high values has negative SHAP values and a relatively negative effect on the model for the low level of chronic stress and a positive impact (positive SHAP values) for class 2. Higher values on the social support scale have a positive impact on class 0 and negative effects on class 2, which means that chronic stress is less likely with strong social support.

Figure 2. SHAP feature plot of the 20 most important features: relative importance of each feature based on the average absolute value of the SHAP values. SHAP: SHapley Additive exPlanations; XGBoost: Extreme Gradient Boosting. *In the past 12 months; **per week.

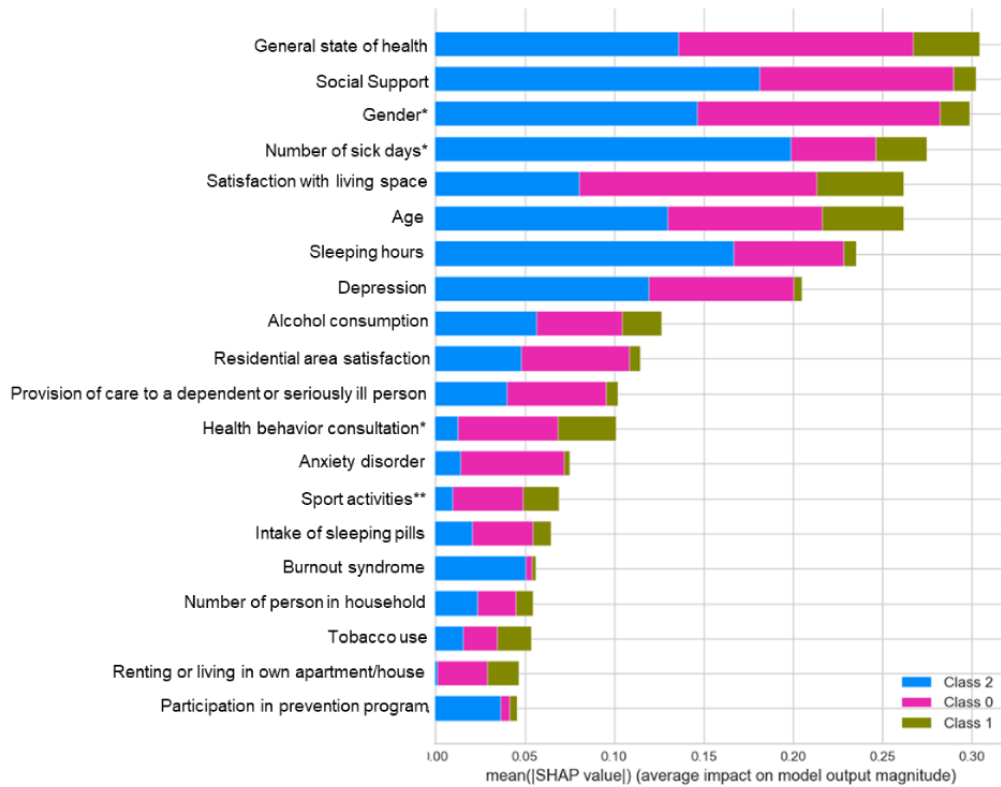


Figure 3. SHAP summary plot. Importance of the representative chronic stress features (top 20) in class 0: each dot is a Shapley value for a particular feature and reflects its impact on a specific class for a given instance, and dots stack up to show density. It is color-coded in accordance with the magnitude to which the value contributes to the model impact (red=high and blue=low). GP: general practitioner; SHAP: SHapley Additive exPlanations. *In the past 12 months; **per week.

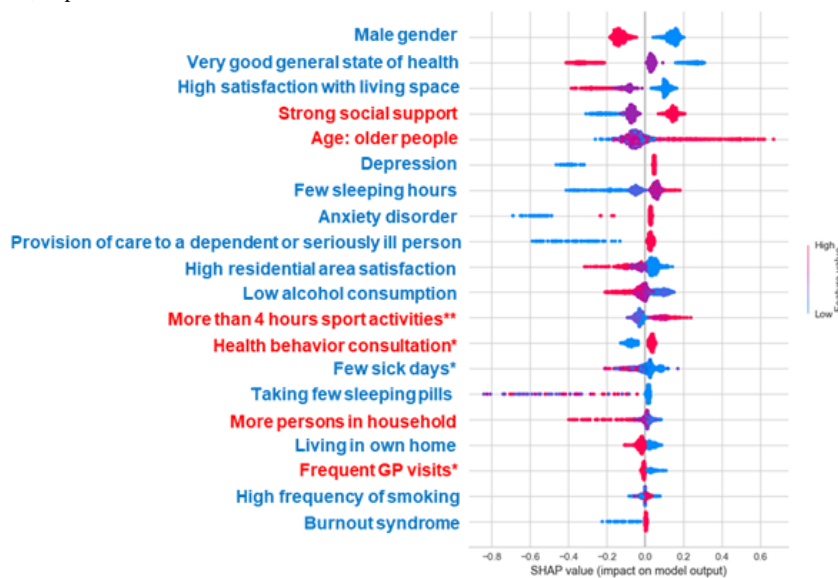
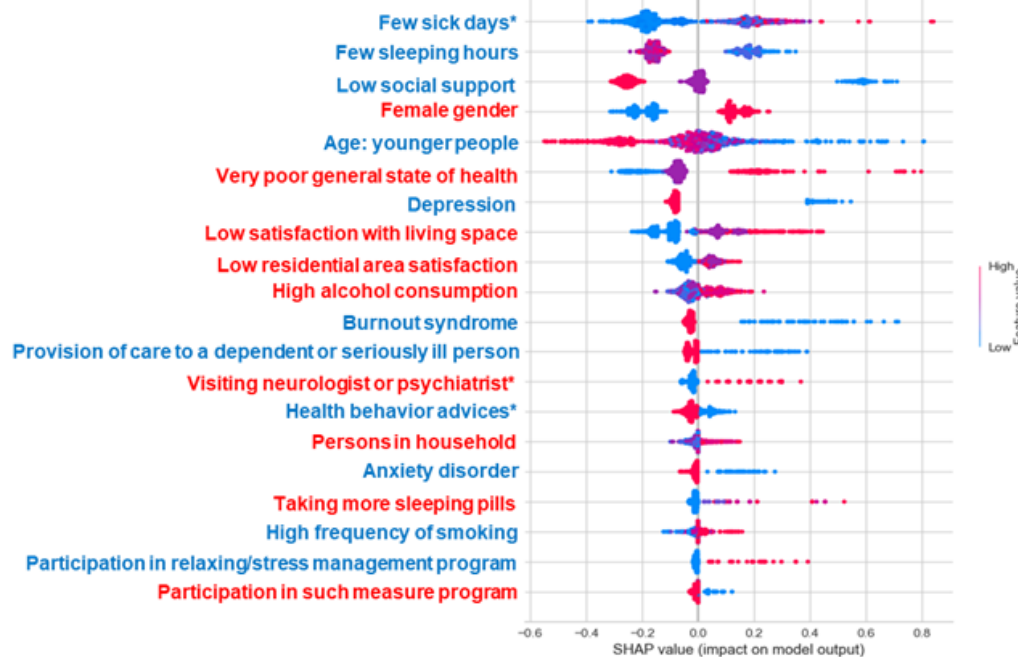


Figure 4. SHAP summary plot. Importance of the representative chronic stress features (top 20) in class 2: each dot is a Shapley value for a particular feature and reflects its impact on a specific class for a given instance, and dots stack up to show density. It is color-coded in accordance with the magnitude to which the value contributes to the model impact (red=high and blue=low). SHAP: SHapley Additive exPlanations. *In the past 12 months.



Discussion

Principal Findings

To our knowledge, this is the first study to select the XGBoost algorithm as an ML multiclass classifier in the prediction of chronic stress as well as the SHAP method to interpret the model's prediction. Based on nationally representative German data, chronic stress was predicted using 34 characteristics of adult participants. We identified male gender, a very good general state of health, high satisfaction with living space, strong social support, enough sleep, and more than 4 hours of sports activities per week as protective factors against chronic stress. These results are in line with those of other studies, which showed that resilience against chronic stress is promoted by social support, family connectedness, and friendship networks in the community [33-36]. For example, with a sample of 24,347 participants from the Canadian General Social Survey, Van der Horst et al [36] determined that good friendship networks are positively associated with less stress, better health, and more social support. A cross-sectional study of 538 nursing students from an Australian university showed that social support positively affect the psychological well-being [37].

Our ML approach allowed for the inclusion of a broad spectrum of individual characteristics, which comprised medical, lifestyle, living space, and social information, while other studies on chronic stress used multivariate models with fewer parameters only. For example, a large cross-sectional study with 34,129 participants from China, Ghana, India, Mexico, Russia, and South Africa showed positive associations of multimorbidity, stroke, depression, and hearing problems with perceived stress

without assessing potential protective factors such as living space and social support [38]. A US cross-sectional telephone survey with 340,847 participants aged between 18 and 85 years documented that psychological well-being, especially stress, improved, but integrated only 5 parameters such as gender, employment status, partnership, and underage children in the household in their model analyzed [39]. In a study with 12,110 working adults from Minnesota, United States, a high level of perceived stress was associated with a higher-fat diet, less exercising, and being a smoker using a multivariate model with 6 variable topics but did not include medical and living circumstances [40].

Strengths and Limitations

This study used the population-based, representative DEGS1 data set, which implies a low risk of selection bias; yet, the results may not be transferrable to other settings. The DEGS1 data, which were collected from 2008 to 2011, may not fully describe current living conditions in Germany, especially the potential effects of the pandemic, which were shown in other studies, were not measured [41]. In our study, the SHAP methodology allowed for a detailed visualization of single feature attributions, which improved the understanding of the ML model.

Conclusions

In this study, we developed an XGBoost ML model to predict chronic stress in adults. The SHAP methodology identified various relevant factors protecting against chronic stress, which need to be considered when developing interventions for stress reduction and improving resilience.

Acknowledgments

We owe special thanks to the Robert Koch Institute, Berlin, Germany, for kindly providing the data set and additional information on the DEGS1 survey.

Conflicts of Interest

None declared.

Multimedia Appendix 1

Hyperparameter Tuning for XGBoost.

[[PDF File \(Adobe PDF File\), 530 KB - ai_v2i1e41868_app1.pdf](#)]

References

1. Marin M, Lord C, Andrews J, Juster R, Sindi S, Arseneault-Lapierre G, et al. Chronic stress, cognitive functioning and mental health. *Neurobiol Learn Mem* 2011 Nov;96(4):583-595. [doi: [10.1016/j.nlm.2011.02.016](https://doi.org/10.1016/j.nlm.2011.02.016)] [Medline: [21376129](https://pubmed.ncbi.nlm.nih.gov/21376129/)]
2. Cohen S, Janicki-Deverts D, Miller GE. Psychological stress and disease. *JAMA* 2007 Oct 10;298(14):1685-1687. [doi: [10.1001/jama.298.14.1685](https://doi.org/10.1001/jama.298.14.1685)] [Medline: [17925521](https://pubmed.ncbi.nlm.nih.gov/17925521/)]
3. Kivimäki M, Steptoe A. Effects of stress on the development and progression of cardiovascular disease. *Nat Rev Cardiol* 2018 Apr 7;15(4):215-229. [doi: [10.1038/nrcardio.2017.189](https://doi.org/10.1038/nrcardio.2017.189)] [Medline: [29213140](https://pubmed.ncbi.nlm.nih.gov/29213140/)]
4. Marcovecchio ML, Chiarelli F. The effects of acute and chronic stress on diabetes control. *Sci Signal* 2012 Oct 23;5(247):pt10. [doi: [10.1126/scisignal.2003508](https://doi.org/10.1126/scisignal.2003508)] [Medline: [23092890](https://pubmed.ncbi.nlm.nih.gov/23092890/)]
5. Landeo-Gutierrez J, Celedón JC. Chronic stress and asthma in adolescents. *Ann Allergy Asthma Immunol* 2020 Oct;125(4):393-398 [FREE Full text] [doi: [10.1016/j.anai.2020.07.001](https://doi.org/10.1016/j.anai.2020.07.001)] [Medline: [32653405](https://pubmed.ncbi.nlm.nih.gov/32653405/)]
6. Hapke U, Maske UE, Scheidt-Nave C, Bode L, Schlack R, Busch MA. [Chronic stress among adults in Germany: results of the German Health Interview and Examination Survey for Adults (DEGS1)]. *Bundesgesundheitsblatt Gesundheitsforschung Gesundheitsschutz* 2013 May;56(5-6):749-754. [doi: [10.1007/s00103-013-1690-9](https://doi.org/10.1007/s00103-013-1690-9)] [Medline: [23703494](https://pubmed.ncbi.nlm.nih.gov/23703494/)]
7. Viehmann A, Kersting C, Thielmann A, Weltermann B. Prevalence of chronic stress in general practitioners and practice assistants: personal, practice and regional characteristics. *PLoS One* 2017;12(5):e0176658 [FREE Full text] [doi: [10.1371/journal.pone.0176658](https://doi.org/10.1371/journal.pone.0176658)] [Medline: [28489939](https://pubmed.ncbi.nlm.nih.gov/28489939/)]
8. Schulz P, Schlotz W, Becker P. *Trierer Inventar zum Chronischen Stress (TICS) Trier Inventory for Chronic Stress (TICS)*. Newburyport, MA: Hogrefe Publishing Corp; 2004.
9. Southwick SM, Bonanno GA, Masten AS, Panter-Brick C, Yehuda R. Resilience definitions, theory, and challenges: interdisciplinary perspectives. *Eur J Psychotraumatol* 2014;5 [FREE Full text] [doi: [10.3402/ejpt.v5.25338](https://doi.org/10.3402/ejpt.v5.25338)] [Medline: [25317257](https://pubmed.ncbi.nlm.nih.gov/25317257/)]
10. Tugade MM, Fredrickson BL. Resilient individuals use positive emotions to bounce back from negative emotional experiences. *J Pers Soc Psychol* 2004 Feb;86(2):320-333 [FREE Full text] [doi: [10.1037/0022-3514.86.2.320](https://doi.org/10.1037/0022-3514.86.2.320)] [Medline: [14769087](https://pubmed.ncbi.nlm.nih.gov/14769087/)]
11. Holz NE, Tost H, Meyer-Lindenberg A. Resilience and the brain: a key role for regulatory circuits linked to social stress and support. *Mol Psychiatry* 2019 Oct 18;25(2):379-396. [doi: [10.1038/s41380-019-0551-9](https://doi.org/10.1038/s41380-019-0551-9)]
12. Schetter CD, Dolbier C. Resilience in the context of chronic stress and health in adults. *Soc Personal Psychol Compass* 2011 Sep;5(9):634-652 [FREE Full text] [doi: [10.1111/j.1751-9004.2011.00379.x](https://doi.org/10.1111/j.1751-9004.2011.00379.x)] [Medline: [26161137](https://pubmed.ncbi.nlm.nih.gov/26161137/)]
13. Henry MS, Gendron L, Tremblay M, Drolet G. Enkephalins: endogenous analgesics with an emerging role in stress resilience. *Neural Plast* 2017;2017:1546125 [FREE Full text] [doi: [10.1155/2017/1546125](https://doi.org/10.1155/2017/1546125)] [Medline: [28781901](https://pubmed.ncbi.nlm.nih.gov/28781901/)]
14. Alpaydin E. *Machine learning*. Cambridge, MA: The MIT Press; 2021.
15. Bonaccorso G. *Machine learning algorithms (first edition)*. Birmingham: Packt Publishing Limited; 2017.
16. Sidey-Gibbons JAM, Sidey-Gibbons CJ. Machine learning in medicine: a practical introduction. *BMC Med Res Methodol* 2019 Mar 19;19(1):64 [FREE Full text] [doi: [10.1186/s12874-019-0681-4](https://doi.org/10.1186/s12874-019-0681-4)] [Medline: [30890124](https://pubmed.ncbi.nlm.nih.gov/30890124/)]
17. Deo RC. Machine learning in medicine. *Circulation* 2015 Nov 17;132(20):1920-1930 [FREE Full text] [doi: [10.1161/CIRCULATIONAHA.115.001593](https://doi.org/10.1161/CIRCULATIONAHA.115.001593)] [Medline: [26572668](https://pubmed.ncbi.nlm.nih.gov/26572668/)]
18. Bozorgmehr A, Thielmann A, Weltermann B. Chronic stress in practice assistants: An analytic approach comparing four machine learning classifiers with a standard logistic regression model. *PLoS One* 2021;16(5):e0250842 [FREE Full text] [doi: [10.1371/journal.pone.0250842](https://doi.org/10.1371/journal.pone.0250842)] [Medline: [33945572](https://pubmed.ncbi.nlm.nih.gov/33945572/)]
19. Gößwald A, Lange M, Dölle R, Hölling H. [The first wave of the German Health Interview and Examination Survey for Adults (DEGS1): participant recruitment, fieldwork, and quality management]. *Bundesgesundheitsblatt Gesundheitsforschung Gesundheitsschutz* 2013 May 25;56(5-6):611-619. [doi: [10.1007/s00103-013-1671-z](https://doi.org/10.1007/s00103-013-1671-z)] [Medline: [23703477](https://pubmed.ncbi.nlm.nih.gov/23703477/)]
20. Beise U. Prävention und Gesundheitsförderung. In: Beise U, Heimes S, Schwarz W, editors. *Gesundheits- und Krankheitslehre*. Berlin: Springer; 2013:27-34.

21. Petrowski K, Paul S, Albani C, Brähler E. Factor structure and psychometric properties of the trier inventory for chronic stress (TICS) in a representative German sample. *BMC Med Res Methodol* 2012 Apr 01;12:42 [[FREE Full text](#)] [doi: [10.1186/1471-2288-12-42](https://doi.org/10.1186/1471-2288-12-42)] [Medline: [22463771](https://pubmed.ncbi.nlm.nih.gov/22463771/)]
22. Schulz P, Schlotz W. Trierer Inventar zur Erfassung von chronischem Streß (TICS): Skalenkonstruktion, teststatistische Überprüfung und Validierung der Skala Arbeitsüberlastung. *Diagnostica* 1999 Jan;45(1):8-19. [doi: [10.1026/0012-1924.45.1.8](https://doi.org/10.1026/0012-1924.45.1.8)]
23. Borkin D, Némethová A, Micháková G, Maiorov K. Impact of data normalization on classification model accuracy. *Research Papers Faculty of Materials Science and Technology Slovak University of Technology* 2019;27(45):79-84 [[FREE Full text](#)] [doi: [10.2478/rput-2019-0029](https://doi.org/10.2478/rput-2019-0029)]
24. Murti DMP, Pujianto U, Wibawa AP, Akbar MI. K-nearest neighbor (K-NN) based missing data imputation. 2019 Presented at: 5th International Conference on Science in Information Technology (ICSITech); October 23-24, 2019; Yogyakarta, Indonesia. [doi: [10.1109/icsitech46713.2019.8987530](https://doi.org/10.1109/icsitech46713.2019.8987530)]
25. Elreedy D, Atiya AF. A comprehensive analysis of synthetic minority oversampling technique (SMOTE) for handling class imbalance. *Information Sciences* 2019 Dec;505:32-64. [doi: [10.1016/j.ins.2019.07.070](https://doi.org/10.1016/j.ins.2019.07.070)]
26. Verhaeghe J, Van Der Donck J, Ongenaes F, Van Hoecke S. Powershap: a power-full Shapley feature selection method. arXiv. Preprint posted online June 16, 2022 2022 [[FREE Full text](#)] [doi: [10.1007/978-3-031-26387-3_5](https://doi.org/10.1007/978-3-031-26387-3_5)]
27. Chen T, Guestrin C. XGBoost: a scalable tree boosting system. 2016 Presented at: 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining; August 13-17, 2016; San Francisco, CA. [doi: [10.1145/2939672.2939785](https://doi.org/10.1145/2939672.2939785)]
28. Che D, Liu Q, Rasheed K, Tao X. Decision tree and ensemble learning algorithms with their applications in bioinformatics. *Adv Exp Med Biol* 2011;696:191-199. [doi: [10.1007/978-1-4419-7046-6_19](https://doi.org/10.1007/978-1-4419-7046-6_19)] [Medline: [21431559](https://pubmed.ncbi.nlm.nih.gov/21431559/)]
29. Grandini M, Bagli E, Visani G. Metrics for multi-class classification: an overview. arXiv. Preprint posted online August 13, 2020 2020 [[FREE Full text](#)]
30. Lundberg S, Lee SI. A unified approach to interpreting model predictions. arXiv. Preprint posted online May 22, 2017 2017 [[FREE Full text](#)]
31. Lundberg SM, Erion G, Chen H, DeGrave A, Prutkin JM, Nair B, et al. From local explanations to global understanding with explainable AI for trees. *Nat Mach Intell* 2020 Jan;2(1):56-67 [[FREE Full text](#)] [doi: [10.1038/s42256-019-0138-9](https://doi.org/10.1038/s42256-019-0138-9)] [Medline: [32607472](https://pubmed.ncbi.nlm.nih.gov/32607472/)]
32. Winter E. The Shapley value. In: *Handbook of Game Theory with Economic Applications*. Amsterdam: Elsevier; 2002:2025-2054.
33. Taylor SE. Social support: a review. In: Friedman HS, editor. *The Oxford handbook of health psychology*. Oxford: Oxford University Press; 2011:189-214.
34. Lepore SJ, Evans GW, Schneider ML. Dynamic role of social support in the link between chronic stress and psychological distress. *J Pers Soc Psychol* 1991 Dec;61(6):899-909. [doi: [10.1037//0022-3514.61.6.899](https://doi.org/10.1037//0022-3514.61.6.899)] [Medline: [1774628](https://pubmed.ncbi.nlm.nih.gov/1774628/)]
35. Thomas PA, Liu H, Umberson D. Family relationships and well-being. *Innov Aging* 2017 Nov;1(3):igx025 [[FREE Full text](#)] [doi: [10.1093/geroni/igx025](https://doi.org/10.1093/geroni/igx025)] [Medline: [29795792](https://pubmed.ncbi.nlm.nih.gov/29795792/)]
36. van der Horst M, Coffé H. How friendship network characteristics influence subjective well-being. *Soc Indic Res* 2012 Jul;107(3):509-529 [[FREE Full text](#)] [doi: [10.1007/s11205-011-9861-2](https://doi.org/10.1007/s11205-011-9861-2)] [Medline: [22707845](https://pubmed.ncbi.nlm.nih.gov/22707845/)]
37. He FX, Turnbull B, Kirshbaum MN, Phillips B, Klainin-Yobas P. Assessing stress, protective factors and psychological well-being among undergraduate nursing students. *Nurse Educ Today* 2018 Sep;68:4-12. [doi: [10.1016/j.nedt.2018.05.013](https://doi.org/10.1016/j.nedt.2018.05.013)] [Medline: [29870871](https://pubmed.ncbi.nlm.nih.gov/29870871/)]
38. Stubbs B, Vancampfort D, Veronese N, Schofield P, Lin P, Tseng P, et al. Multimorbidity and perceived stress: a population-based cross-sectional study among older adults across six low- and middle-income countries. *Maturitas* 2018 Jan;107:84-91 [[FREE Full text](#)] [doi: [10.1016/j.maturitas.2017.10.007](https://doi.org/10.1016/j.maturitas.2017.10.007)] [Medline: [29169587](https://pubmed.ncbi.nlm.nih.gov/29169587/)]
39. Stone AA, Schwartz JE, Broderick JE, Deaton A. A snapshot of the age distribution of psychological well-being in the United States. *Proc Natl Acad Sci U S A* 2010 Jun 01;107(22):9985-9990 [[FREE Full text](#)] [doi: [10.1073/pnas.1003744107](https://doi.org/10.1073/pnas.1003744107)] [Medline: [20479218](https://pubmed.ncbi.nlm.nih.gov/20479218/)]
40. Ng DM, Jeffery RW. Relationships between perceived stress and health behaviors in a sample of working adults. *Health Psychol* 2003 Nov;22(6):638-642. [doi: [10.1037/0278-6133.22.6.638](https://doi.org/10.1037/0278-6133.22.6.638)] [Medline: [14640862](https://pubmed.ncbi.nlm.nih.gov/14640862/)]
41. Schelhorn I, Ecker A, Lütke MN, Rehm S, Tran T, Bereznaï JL, et al. Psychological burden during the COVID-19 pandemic in Germany. *Front Psychol* 2021;12:640518 [[FREE Full text](#)] [doi: [10.3389/fpsyg.2021.640518](https://doi.org/10.3389/fpsyg.2021.640518)] [Medline: [34557124](https://pubmed.ncbi.nlm.nih.gov/34557124/)]

Abbreviations

- AUC:** area under the receiver operating characteristic curve
- DEGS1:** German Health Interview and Examination Survey for Adults
- GP:** general practitioner
- KNN:** K-nearest neighbors
- ML:** machine learning

PrA: practice assistant

ROC: receiver operating characteristic

SHAP: SHapley Additive exPlanations

TICS-SSCS: Screening Scale of the Trier Inventory for the Assessment of Chronic Stress

XGBoost: Extreme Gradient Boosting

Edited by K El Emam; submitted 12.08.22; peer-reviewed by W Klement, J Li; comments to author 14.11.22; revised version received 06.01.23; accepted 03.03.23; published 16.05.23.

Please cite as:

Bozorgmehr A, Weltermann B

Prediction of Chronic Stress and Protective Factors in Adults: Development of an Interpretable Prediction Model Based on XGBoost and SHAP Using National Cross-sectional DEGS1 Data

JMIR AI 2023;2:e41868

URL: <https://ai.jmir.org/2023/1/e41868>

doi: [10.2196/41868](https://doi.org/10.2196/41868)

PMID: [38875576](https://pubmed.ncbi.nlm.nih.gov/38875576/)

©Arezoo Bozorgmehr, Birgitta Weltermann. Originally published in JMIR AI (<https://ai.jmir.org>), 16.05.2023. This is an open-access article distributed under the terms of the Creative Commons Attribution License (<https://creativecommons.org/licenses/by/4.0/>), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR AI, is properly cited. The complete bibliographic information, a link to the original publication on <https://www.ai.jmir.org/>, as well as this copyright and license information must be included.

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

<https://www.jmirpublications.com/>