Contents

Editorial

Reporting and Methodological Observations on Prognostic and Diagnostic Machine Learning Studies (e47995)
Khaled El Emam, William Klement, Bradley Malin. ................................................................. 4

Reviews

Strategies to Improve the Impact of Artificial Intelligence on Health Equity: Scoping Review (e42936)
Carl Berdahl, Lawrence Baker, Sean Mann, Osonde Osoba, Federico Girosi. ................................. 8

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. ................................................................. 33

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. .................................................................................................. 287

Original Papers

Forecasting Artificial Intelligence Trends in Health Care: Systematic International Patent Analysis (e47283)
Stan Benjamens, Pranav Singh Dhunnoo, Mártón Görög, Bertalan Mesko. ................................. 23

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. ................................................................. 46

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. ................................................................. 60

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. ................................................................. 73
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

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, Didac Cumelles, Júlia Rodríguez-Comas, Javier García López, Mª Fernández Planas

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

Patient Embeddings From Diagnosis Codes for Health Care Prediction Tasks: Pat2Vec Machine Learning Framework (e40755)  
Edgar Steiger, Lars Kroll

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

Application of Artificial Intelligence to the Monitoring of Medication Adherence for Tuberculosis Treatment in Africa: Algorithm Development and Validation (e40167)  
Juliet Sekandi, Wei Li, Ronghang Zhu, Patrick Kagwara, Ernest Mwebaze, Sheng Li

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

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

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

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

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

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 Nasari, Sonia Skamene, Marwan Tolba, Mame Faye, Paul Ramia, Julia Khrguijan, Marc David, John Kildea
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 Al, Parag Aggarwal, Qi Pan, William Oh, Gustavo Stolovitzky, Eric Schadt, Xiaoyan Wang. ........................................ 250

A Trainable Open-Source Machine Learning Accelerometer Activity Recognition Toolbox: Deep Learning Approach (e42337)
Fluri Wieland, Claudio Nigg. .......................................................... 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. ................................. 276

Predicting Treatment Interruption Among People Living With HIV in Nigeria: Machine Learning Approach (e44432)
Matthew-David Ogbechie, Christa Fischer Walker, Mu-Tien Lee, Amina Abba Gana, Abimbola Oduola, Augustine Idemudia, Matthew Edor, Emily Harris, Jessica Stephens, Xiaoming Gao, Pai-Lien Chen, Navindra Persaud. .................................................. 300

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. ................................................................. 309

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. ................................................................. 327

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. ................................. 178
Editorial

Reporting and Methodological Observations on Prognostic and Diagnostic Machine Learning Studies

Khaled El Emam¹ ², BEng, PhD; William Klement¹ 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

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**


**Abbreviations**

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

Carl Thomas Berdahl1,2,3, 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

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... Pursing 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 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, supervised 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% (120/1987) 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.

<table>
<thead>
<tr>
<th>Data sets, variable selection, and health equity</th>
</tr>
</thead>
<tbody>
<tr>
<td>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.</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Limitations in evaluating equity-related outcomes</th>
</tr>
</thead>
<tbody>
<tr>
<td>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.</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Availability of equity-related information on AI algorithm performance</th>
</tr>
</thead>
<tbody>
<tr>
<td>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.</td>
</tr>
</tbody>
</table>

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 κ=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].
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\textsuperscript{a} and health equity that were abstracted from the literature.

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

\textsuperscript{a}AI: artificial intelligence.
Table 2. Strategies to address AI equity issues that were abstracted from the literature.

<table>
<thead>
<tr>
<th>Category and strategy</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Background Context</strong></td>
<td></td>
</tr>
<tr>
<td>Foster diversity</td>
<td>Create AI development teams with diverse characteristics, experiences, and roles to increase consideration of equity throughout development and decrease blind spots.</td>
</tr>
<tr>
<td>Train developers and users</td>
<td>Train AI developers and users in equity considerations and the ethical implications of AI, as these topics may be unfamiliar to some.</td>
</tr>
<tr>
<td>Engage the broader community</td>
<td>Foster community involvement throughout development, from conception to postdeployment, to increase the likelihood that developers prioritize equity concerns.</td>
</tr>
<tr>
<td>Improve governance</td>
<td>Enact robust regulation and industry standards to align AI applications with social norms, including equity, safety, and transparency.</td>
</tr>
<tr>
<td><strong>Data Characteristics</strong></td>
<td></td>
</tr>
<tr>
<td>Improve diversity, quality, or quantity of data</td>
<td>Train models with large, diverse samples that are representative of the target population for the application and contain all relevant features.</td>
</tr>
<tr>
<td>Exclude sensitive variables to correct for bias</td>
<td>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.</td>
</tr>
<tr>
<td>Include sensitive variables to correct for bias</td>
<td>Include sensitive variables to improve model accuracy, increase explanatory power, and enable easier testing for inequitable impact.</td>
</tr>
<tr>
<td><strong>Model Design</strong></td>
<td></td>
</tr>
<tr>
<td>Enforce fairness goals</td>
<td>Formulate a fairness norm and enforce it in the model by editing the input data, objective function, or model outputs.</td>
</tr>
<tr>
<td>Improve interpretability or explainability of the algorithm</td>
<td>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.</td>
</tr>
<tr>
<td>Evaluate disparities in model performance</td>
<td>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.</td>
</tr>
<tr>
<td>Use equity-focused checklists, guidelines, and similar tools</td>
<td>Incorporate equity-focused checklists into workflows for developers, reviewers of AI models, healthcare providers using an application, or patients who want to understand algorithm outputs.</td>
</tr>
<tr>
<td><strong>Deployment Practices</strong></td>
<td></td>
</tr>
<tr>
<td>Increase model reporting and transparency</td>
<td>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.</td>
</tr>
<tr>
<td>Seek or provide restitution for those negatively impacted by AI</td>
<td>Proactively provide restitution to those harmed by AI or create legal frameworks so they can seek restitution.</td>
</tr>
<tr>
<td>Avoid or reduce use of AI</td>
<td>Consider discontinuing model use if equity sequelae are severe or if improvement efforts have been fruitless.</td>
</tr>
<tr>
<td>Provide resources to those with less access to AI</td>
<td>Improve access to AI for disadvantaged groups and low-income countries by subsidizing infrastructure, creating education programs, or hosting AI conferences in these locations.</td>
</tr>
</tbody>
</table>

*AI: 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.

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

<sup>a</sup>AI: artificial intelligence.

<sup>b</sup>Only 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. ID9QT0822-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


46. Chouldechova A, Roth A. The frontiers of fairness in machine learning. arXiv. Preprint posted online on October 20, 2018 [FREE Full text]


### Abbreviations

<table>
<thead>
<tr>
<th>Abbreviation</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>AI</td>
<td>artificial intelligence</td>
</tr>
<tr>
<td>BeHEMoTh</td>
<td>Behavior, Health condition, Exclusions, Models, or Theories</td>
</tr>
<tr>
<td>FDA</td>
<td>Food and Drug Administration</td>
</tr>
<tr>
<td>IEEE</td>
<td>Institute of Electrical and Electronics Engineers</td>
</tr>
<tr>
<td>PRISMA-ScR</td>
<td>Preferred Reporting Items for Systematic Reviews and Meta-Analyses Extension for Scoping Reviews</td>
</tr>
<tr>
<td>PROBAST</td>
<td>Prediction Model Risk of Bias Assessment Tool</td>
</tr>
</tbody>
</table>

©Carl Thomas Berdahl, Lawrence Baker, Sean Mann, Osonde Osoba, Federico Girosi. 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.
Forecasting Artificial Intelligence Trends in Health Care: Systematic International Patent Analysis

Stan Benjamens¹, MD, PhD; PranavSingh Dhunnoo²,³, 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.


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

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 electrocardiograph 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 appeared 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 intervals.

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.

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

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

The top 10 countries from where patents were filed were as follows:

1. China
2. United States
3. South Korea
4. Germany
5. Japan
6. The Netherlands
7. Canada
8. India
9. United Kingdom
10. France
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 downturns, 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.

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

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

---

**References**


5. Five IP Offices. URL: https://www.fiveipoffices.org/home [accessed 2023-05-10]


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]


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

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 translatable into real COVID-19 health care environments.

doi: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 framework.

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

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 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.
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<0.003 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 (κ=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.
**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; \( P = 2.39 \times 10^{-05} \)) 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.
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 0.31; proprietary software studies mean score 0.75, SE 0.14; P=.43), and “nonmaleficence” (open source software studies mean score 0.82, SE 0.22; proprietary software studies mean score 0.70, SE 0.35; P=.74).
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 privacy, 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=0.04; 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=0.03) and “internal validity” (high-income study mean score 1.58, SE 0.08; medium-income study mean score 1.88, SE 0.08; P=0.04) 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=0.02), “safety and quality” (imaging study mean score 0.16, SE 0.05; nonimaging study mean score 0.36, SE 0.14; P=0.15), “nonmaleficence” (imaging study mean score 0.04, SE 0.02; nonimaging study mean score 0.18, SE 0.07; P=0.05), and “number of services” (imaging study mean score 0.25, SE 0.06; nonimaging study mean score 0.55, SE 0.11; P=0.02) 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 translatable. 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 rigorous 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


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

https://ai.jmir.org/2023/1/e42313

JMI RAI 2023 | vol. 2 | e42313 | p.44
(page number not for citation purposes)
Deep Learning Transformer Models for Building a Comprehensive and Real-time Trauma Observatory: Development and Validation Study

Gabrielle Chenais, MMid, MScPH, MPHd; Cédric Gil-Jardiné, MD, PhD; Hélène Touchais, MCPM; Marta Avalos Fernández, PhD; Benjamin Contrand, MSc; Eric Tellier, MD; Xavier Combes, MD; Loïck Bourdois, MSc; Philippe Revel, MD; Emmanuel Lagarde, PhD

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 and a 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 F₁-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.

doi:10.2196/40843

KEYWORDS

deep learning; public health; trauma; emergencies; natural language processing; transformers

Corresponding Author:

Gabrielle Chenais, MMid, MScPH, MPHd
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
**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 (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...
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). BiGPT2 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, Cosson 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 Iev 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 Injuries,” 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 70,999, 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 (e.g., “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).

**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 F1-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 1. Transformer models’ sizes and configurations.

<table>
<thead>
<tr>
<th>Model</th>
<th>Layers</th>
<th>Attention heads</th>
<th>Embedding dimension</th>
<th>Parameters (millions)</th>
<th>Pretraining corpus size (GB)</th>
</tr>
</thead>
<tbody>
<tr>
<td>CamemBERT-base-CCNETa</td>
<td>12</td>
<td>12</td>
<td>768</td>
<td>110</td>
<td>135</td>
</tr>
<tr>
<td>FlauBERT-base-cased</td>
<td>12</td>
<td>12</td>
<td>768</td>
<td>138</td>
<td>71</td>
</tr>
<tr>
<td>BelGPT2</td>
<td>12</td>
<td>12</td>
<td>768</td>
<td>117</td>
<td>57.9</td>
</tr>
<tr>
<td>GPTanam</td>
<td>12</td>
<td>12</td>
<td>768</td>
<td>117</td>
<td>58.6</td>
</tr>
</tbody>
</table>

aCCNET: criss-cross attention for semantic segmentation.
Table 2. Label distribution among train, validation, and test data sets.

<table>
<thead>
<tr>
<th>Type of trauma</th>
<th>Train data set (n=14,532, 64.64%), n (%)</th>
<th>Validation data set (n=3634, 16.16%), n (%)</th>
<th>Test data set (n=4315, 19.19%), n (%)</th>
<th>Total (N=22,481, 100%), n (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Accident of exposure to bodily fluids</td>
<td>132 (0.91)</td>
<td>40 (1.1)</td>
<td>41 (1)</td>
<td>213 (0.9)</td>
</tr>
<tr>
<td>Assault</td>
<td>1587 (10.92)</td>
<td>393 (10.81)</td>
<td>498 (11.54)</td>
<td>2478 (11.02)</td>
</tr>
<tr>
<td>MVA a</td>
<td>2028 (13.95)</td>
<td>495 (13.62)</td>
<td>568 (13.16)</td>
<td>3091 (13.75)</td>
</tr>
<tr>
<td>Foreign body in eye</td>
<td>642 (4.42)</td>
<td>180 (5)</td>
<td>186 (4.31)</td>
<td>1008 (4.48)</td>
</tr>
<tr>
<td>Fall</td>
<td>4778 (32.87)</td>
<td>1162 (31.97)</td>
<td>1554 (36.01)</td>
<td>7494 (33.33)</td>
</tr>
<tr>
<td>Sport accident</td>
<td>1311 (9)</td>
<td>341 (9.38)</td>
<td>371 (8.59)</td>
<td>2023 (9)</td>
</tr>
<tr>
<td>Intentional injury</td>
<td>341 (2.34)</td>
<td>73 (2)</td>
<td>112 (2.59)</td>
<td>526 (2.33)</td>
</tr>
<tr>
<td>Other trauma</td>
<td>3713 (25.55)</td>
<td>950 (26.14)</td>
<td>985 (22.82)</td>
<td>5648 (25.12)</td>
</tr>
</tbody>
</table>

aMVA: motor vehicle accident.

Table 3. Train, validation, and test data set characteristics.

<table>
<thead>
<tr>
<th>Year of ED b venue, n (%)</th>
<th>Train data set (n=14,532)</th>
<th>Validation data set (n=3634)</th>
<th>Test data set (n=4315)</th>
<th>Total (N=22,481)</th>
</tr>
</thead>
<tbody>
<tr>
<td>2012</td>
<td>218 (1.5)</td>
<td>52 (1.43)</td>
<td>66 (1.52)</td>
<td>336 (1.49)</td>
</tr>
<tr>
<td>2013</td>
<td>1389 (12.2)</td>
<td>359 (12.4)</td>
<td>418 (12.3)</td>
<td>2166 (12.2)</td>
</tr>
<tr>
<td>2014</td>
<td>1444 (12.6)</td>
<td>385 (13.3)</td>
<td>386 (11.3)</td>
<td>2215 (12.3)</td>
</tr>
<tr>
<td>2015</td>
<td>1502 (13.1)</td>
<td>326 (11.2)</td>
<td>425 (12.5)</td>
<td>2253 (12.6)</td>
</tr>
<tr>
<td>2016</td>
<td>1419 (12.4)</td>
<td>365 (12.6)</td>
<td>426 (12.6)</td>
<td>2210 (12.3)</td>
</tr>
<tr>
<td>2017</td>
<td>1493 (13.1)</td>
<td>370 (12.8)</td>
<td>461 (13.5)</td>
<td>2324 (12.9)</td>
</tr>
<tr>
<td>2018</td>
<td>1425 (12.5)</td>
<td>405 (13.9)</td>
<td>474 (13.9)</td>
<td>2304 (13.5)</td>
</tr>
<tr>
<td>2019</td>
<td>690 (6)</td>
<td>175 (6)</td>
<td>218 (6.4)</td>
<td>1083 (6.2)</td>
</tr>
<tr>
<td>2020</td>
<td>1856 (16.2)</td>
<td>468 (16.1)</td>
<td>532 (15.6)</td>
<td>2856 (16)</td>
</tr>
<tr>
<td>Missing values</td>
<td>3118 (27.3)</td>
<td>737 (25.4)</td>
<td>899 (26.4)</td>
<td>4724 (20.9)</td>
</tr>
</tbody>
</table>

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.
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.**

<table>
<thead>
<tr>
<th>Type of trauma</th>
<th>Test data set</th>
<th>Micro $F_1$-scores</th>
<th>TF-IDF/SVM$^b$</th>
<th>CamemBERT</th>
<th>FlauBERT</th>
<th>BelGPT2</th>
<th>GPTanam</th>
</tr>
</thead>
<tbody>
<tr>
<td>Accident of exposure to bodily fluids</td>
<td>41 (1)</td>
<td>0.83</td>
<td>0.84</td>
<td>0.84</td>
<td>0.83</td>
<td>0.91 $^c$</td>
<td></td>
</tr>
<tr>
<td>Assault</td>
<td>498 (11.54)</td>
<td>0.91</td>
<td>0.91</td>
<td>0.92</td>
<td>0.91</td>
<td>0.96</td>
<td></td>
</tr>
<tr>
<td>MVA$^d$</td>
<td>568 (13.16)</td>
<td>0.91</td>
<td>0.90</td>
<td>0.91</td>
<td>0.91</td>
<td>0.97</td>
<td></td>
</tr>
<tr>
<td>Foreign body in eye</td>
<td>186 (4.3)</td>
<td>0.79</td>
<td>0.84</td>
<td>0.82</td>
<td>0.82</td>
<td>0.97</td>
<td></td>
</tr>
<tr>
<td>Fall</td>
<td>1554 (36.01)</td>
<td>0.9</td>
<td>0.92</td>
<td>0.91</td>
<td>0.92</td>
<td>0.98</td>
<td></td>
</tr>
<tr>
<td>Sport accident</td>
<td>371 (8.6)</td>
<td>0.82</td>
<td>0.83</td>
<td>0.83</td>
<td>0.85</td>
<td>0.94</td>
<td></td>
</tr>
<tr>
<td>Intentional injury</td>
<td>112 (2.6)</td>
<td>0.75</td>
<td>0.76</td>
<td>0.73</td>
<td>0.77</td>
<td>0.84</td>
<td></td>
</tr>
<tr>
<td>Other trauma</td>
<td>985 (22.8)</td>
<td>0.8</td>
<td>0.83</td>
<td>0.82</td>
<td>0.85</td>
<td>0.98</td>
<td></td>
</tr>
<tr>
<td>Micro $F_1$-score</td>
<td>N/A$^e$</td>
<td>0.864</td>
<td>0.878</td>
<td>0.873</td>
<td>0.887</td>
<td>0.969</td>
<td></td>
</tr>
<tr>
<td>Macro precision</td>
<td>N/A$^e$</td>
<td>0.860</td>
<td>0.880</td>
<td>0.880</td>
<td>0.89</td>
<td>0.970</td>
<td></td>
</tr>
</tbody>
</table>

$^a$TF-IDF: term frequency–inverse document frequency.
$^b$SVM: support vector machine.
$^c$The best $F_1$-scores are in italic.
$^d$MVA: motor vehicle accident.
$^e$N/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 Soubellet, 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 (Délegation 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]


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
TARPO: Traitement Automatique des Résumés de Passage aux urgences dans le but de créer un Observatoire National
**TF-IDF:** term frequency–inverse document frequency
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

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 care 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

https://ai.jmir.org/2023/1/e42253

Bowers et al

JMIR AI 2023 | vol. 2 | e42253 | p.61

(page number not for citation purposes)
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. Models were generated using a training cohort (255,020/318,774, 80% of the sample) to validate and assess generalization to new cases. 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 informed from claims, EHRs, and administrative member records.

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 informed 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=.92; 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.

<table>
<thead>
<tr>
<th>Characteristic</th>
<th>Total sample (n=318,774)</th>
<th>Alive (n=306,227, 96.1%)</th>
<th>Deceased (n=12,547, 3.9%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gender, n (%)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Female</td>
<td>181,158 (56.8)</td>
<td>174,640 (57.0)</td>
<td>6518 (51.9)</td>
</tr>
<tr>
<td>Male</td>
<td>136,970 (43.0)</td>
<td>130,941 (42.8)</td>
<td>6092 (48.1)</td>
</tr>
<tr>
<td>Missing/not available</td>
<td>646 (0.2)</td>
<td>646 (0.2)</td>
<td>0 (0)</td>
</tr>
<tr>
<td>Age (years), mean (SD)</td>
<td>70.7 (11.5)</td>
<td>70.4 (11.5)</td>
<td>77.2 (9.7)</td>
</tr>
<tr>
<td>Medical diagnoses, n (%)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Chronic respiratory disease</td>
<td>56,734 (10.4)</td>
<td>52,183 (10.2)</td>
<td>4551 (14.0)</td>
</tr>
<tr>
<td>Heart failure</td>
<td>54,702 (10.1)</td>
<td>50,254 (9.8)</td>
<td>4448 (13.7)</td>
</tr>
<tr>
<td>Cancer</td>
<td>44,145 (8.1)</td>
<td>40,985 (8.0)</td>
<td>3160 (9.7)</td>
</tr>
<tr>
<td>Stroke</td>
<td>21,338 (3.9)</td>
<td>19,327 (3.8)</td>
<td>2011 (6.2)</td>
</tr>
<tr>
<td>Dementia or Alzheimer disease</td>
<td>15,626 (2.9)</td>
<td>13,018 (2.5)</td>
<td>2608 (8.0)</td>
</tr>
<tr>
<td>Hypertension</td>
<td>204,405 (37.6)</td>
<td>195,035 (38.2)</td>
<td>9370 (28.8)</td>
</tr>
<tr>
<td>Diabetes</td>
<td>146,394 (26.9)</td>
<td>139,999 (27.4)</td>
<td>6395 (19.7)</td>
</tr>
<tr>
<td>Medical service utilization, mean (SD)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total care visits per year(^a)</td>
<td>20.8 (39.5)</td>
<td>20.2 (38.2)</td>
<td>36.7 (60.9)</td>
</tr>
<tr>
<td>Emergency room visits per year</td>
<td>0.4 (1.1)</td>
<td>0.4 (1.1)</td>
<td>0.9 (1.7)</td>
</tr>
<tr>
<td>Pharmacy utilization, mean (SD)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total unique medications prescribed</td>
<td>9.04 (7.4)</td>
<td>8.9 (7.3)</td>
<td>11.7 (8.3)</td>
</tr>
<tr>
<td>Number of prescribed medications per day</td>
<td>8.11 (12.0)</td>
<td>8.0 (12.1)</td>
<td>9.8 (9.9)</td>
</tr>
<tr>
<td>Laboratory utilization, mean (SD)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total unique lab-related CPT(^b) codes</td>
<td>8.7 (8.4)</td>
<td>8.6 (8.2)</td>
<td>11.7 (11.0)</td>
</tr>
<tr>
<td>Social determinants index (SDI)(^c), mean (SD)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Weighted SDI score(^d)</td>
<td>58.41 (8.65)</td>
<td>58.43 (8.67)</td>
<td>58.09 (8.08)</td>
</tr>
<tr>
<td>Unweighted SDI score(^d)</td>
<td>56.94 (10.12)</td>
<td>56.98 (10.13)</td>
<td>55.91 (9.63)</td>
</tr>
</tbody>
</table>

\(^a\)Includes all inpatient and outpatient visits.
\(^c\)Higher is better.
\(^d\)100 points maximum.
Table 2. Model summary and performance comparison (holdout cohort).

<table>
<thead>
<tr>
<th>Measure</th>
<th>Model 1 (M1; baseline)</th>
<th>Model 2 (M2)</th>
<th>Model 3 (M3; final)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total model features, n</td>
<td>7</td>
<td>899</td>
<td>907</td>
</tr>
<tr>
<td>Model input summary</td>
<td>Demographics&lt;sup&gt;a&lt;/sup&gt;, High-risk utilization indicators&lt;sup&gt;b,c&lt;/sup&gt;</td>
<td>Demographics&lt;sup&gt;a&lt;/sup&gt;, High-risk utilization indicators&lt;sup&gt;b,c&lt;/sup&gt;; Medical, lab, and pharmacy utilization&lt;sup&gt;c&lt;/sup&gt;</td>
<td>Demographics&lt;sup&gt;a&lt;/sup&gt;; High-risk utilization indicators&lt;sup&gt;b,c&lt;/sup&gt;; Medical, lab, and pharmacy utilization&lt;sup&gt;c&lt;/sup&gt;; SDI&lt;sup&gt;d&lt;/sup&gt; scores&lt;sup&gt;a&lt;/sup&gt;</td>
</tr>
<tr>
<td>Model performance (holdout cohort)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>AUC&lt;sup&gt;e&lt;/sup&gt; (95% CI)</td>
<td>0.736 (0.728-0.744)</td>
<td>0.834 (0.828-0.840)</td>
<td>0.839 (0.833-0.845)</td>
</tr>
<tr>
<td>True positive rate&lt;sup&gt;f&lt;/sup&gt;</td>
<td>0.105</td>
<td>0.320</td>
<td>0.2993</td>
</tr>
<tr>
<td>PPV&lt;sup&gt;f,g&lt;/sup&gt;</td>
<td>0.212</td>
<td>0.264</td>
<td>0.2991</td>
</tr>
<tr>
<td>False positive rate&lt;sup&gt;d&lt;/sup&gt;</td>
<td>0.016</td>
<td>0.037</td>
<td>0.029</td>
</tr>
<tr>
<td>True negative rate&lt;sup&gt;f&lt;/sup&gt;</td>
<td>0.984</td>
<td>0.963</td>
<td>0.97126</td>
</tr>
<tr>
<td>NPV&lt;sup&gt;f,h&lt;/sup&gt;</td>
<td>0.964</td>
<td>0.972</td>
<td>0.97129</td>
</tr>
<tr>
<td>False negative rate&lt;sup&gt;f&lt;/sup&gt;</td>
<td>0.890</td>
<td>0.679</td>
<td>0.701</td>
</tr>
<tr>
<td>AP&lt;sup&gt;i&lt;/sup&gt;</td>
<td>0.122</td>
<td>0.233</td>
<td>0.243</td>
</tr>
<tr>
<td>Performance comparison (holdout cohort)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Null hypothesis</td>
<td>AUC&lt;sub&gt;M1&lt;/sub&gt; = 0.5</td>
<td>AUC&lt;sub&gt;M2&lt;/sub&gt; – AUC&lt;sub&gt;M1&lt;/sub&gt; = 0.0</td>
<td>AUC&lt;sub&gt;M3&lt;/sub&gt; – AUC&lt;sub&gt;M2&lt;/sub&gt; = 0.0</td>
</tr>
<tr>
<td>z statistic</td>
<td>56.4</td>
<td>19.1</td>
<td>1.2</td>
</tr>
<tr>
<td>P value</td>
<td>&lt;.001</td>
<td>&lt;.001</td>
<td>.19</td>
</tr>
</tbody>
</table>

<sup>a</sup>Source: internal administrative member records.

<sup>b</sup>Source: electronic health record (EHR) data.

<sup>c</sup>Source: claims data.

<sup>d</sup>SDI: social determinants index.

<sup>e</sup>AUC: area under the curve.

<sup>f</sup>Values based on a defined threshold of 0.16.

<sup>g</sup>PPV: positive predictive value.

<sup>h</sup>NPV: negative predictive value.

<sup>i</sup>AP: 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.

(A)

![Receiver Operating Characteristic Curves](image1)

(B)

![Precision Recall Curves](image2)

**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).

<table>
<thead>
<tr>
<th>Feature category and M3 features</th>
<th>M3 ranked importance&lt;sup&gt;a&lt;/sup&gt;</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Demographics (2 inputs)</strong></td>
<td></td>
</tr>
<tr>
<td>Age&lt;sup&gt;b&lt;/sup&gt;</td>
<td>1</td>
</tr>
<tr>
<td>Gender&lt;sup&gt;b&lt;/sup&gt;</td>
<td>2</td>
</tr>
<tr>
<td><strong>Diagnoses (233 inputs)</strong></td>
<td></td>
</tr>
<tr>
<td>Chronic respiratory disease</td>
<td>3</td>
</tr>
<tr>
<td>Kidney disease&lt;sup&gt;b&lt;/sup&gt;</td>
<td>8</td>
</tr>
<tr>
<td>Total patient diagnoses</td>
<td>17</td>
</tr>
<tr>
<td>Dementia</td>
<td>18</td>
</tr>
<tr>
<td><strong>Medical utilization (265 inputs)</strong></td>
<td></td>
</tr>
<tr>
<td>Total patient claims</td>
<td>4</td>
</tr>
<tr>
<td>Average cost of claim</td>
<td>11</td>
</tr>
<tr>
<td>Total CT&lt;sup&gt;c&lt;/sup&gt; scans</td>
<td>13</td>
</tr>
<tr>
<td>Time since last outpatient visit</td>
<td>15</td>
</tr>
<tr>
<td><strong>Pharmacy utilization (198 inputs)</strong></td>
<td></td>
</tr>
<tr>
<td>Antihyperlipidemics</td>
<td>5</td>
</tr>
<tr>
<td>Furosemide</td>
<td>7</td>
</tr>
<tr>
<td>Anti-inflammatory analgesics</td>
<td>9</td>
</tr>
<tr>
<td>Beta blockers</td>
<td>14</td>
</tr>
<tr>
<td>Antidepressants</td>
<td>16</td>
</tr>
<tr>
<td>Diuretics</td>
<td>19</td>
</tr>
<tr>
<td><strong>Laboratory utilization (201 inputs)</strong></td>
<td></td>
</tr>
<tr>
<td>Systemic and topical nasal agents</td>
<td>20</td>
</tr>
<tr>
<td>Lipid panel lab test</td>
<td>6</td>
</tr>
<tr>
<td><strong>Social determinants index (SDI) score (8 inputs)</strong></td>
<td></td>
</tr>
<tr>
<td>Food access</td>
<td>10</td>
</tr>
<tr>
<td>Economy</td>
<td>12</td>
</tr>
</tbody>
</table>

<sup>a</sup>Ranked importance based on positive Shapley Additive Explanations value of features.<br>
<sup>b</sup>M1 feature.<br>
<sup>c</sup>CT: computed tomography.
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


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] [Medline: 23703645]


Abbreviations

AUC: area under the receiver operating characteristic curve
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

©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

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 Charow¹,³, MSc; Inaara Karman⁴,³, BSc; Azra Dhalla¹, MBA; Dalia Al-Mouaswas², HBSc; Jillian Scandiffio¹, MSc; Justin Aling⁴; Mohammad Salhia², MEd; Nadim Lalani³, BA; Scott Overholt³; David Wiljer¹,³,⁶,⁷, 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 synergistic 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.
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 co-creators 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).

<table>
<thead>
<tr>
<th>Characteristic</th>
<th>Value, n (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Demographics</strong></td>
<td></td>
</tr>
<tr>
<td><strong>Age (years)</strong></td>
<td></td>
</tr>
<tr>
<td>Young adult (18-40)</td>
<td>0 (0)</td>
</tr>
<tr>
<td>Middle age (40-60)</td>
<td>7 (58)</td>
</tr>
<tr>
<td>Senior (≥60)</td>
<td>5 (42)</td>
</tr>
<tr>
<td><strong>Sex</strong></td>
<td></td>
</tr>
<tr>
<td>Male</td>
<td>2 (17)</td>
</tr>
<tr>
<td>Female</td>
<td>10 (83)</td>
</tr>
<tr>
<td><strong>Location</strong></td>
<td></td>
</tr>
<tr>
<td>Urban</td>
<td>8 (67)</td>
</tr>
<tr>
<td>Rural</td>
<td>4 (33)</td>
</tr>
<tr>
<td><strong>Comfort with technology</strong></td>
<td></td>
</tr>
<tr>
<td>Not at all comfortable</td>
<td>3 (25)</td>
</tr>
<tr>
<td>Somewhat comfortable</td>
<td>3 (25)</td>
</tr>
<tr>
<td>Very comfortable</td>
<td>6 (50)</td>
</tr>
<tr>
<td><strong>Familiarity with AI\textsuperscript{a}</strong></td>
<td></td>
</tr>
<tr>
<td>Not at all familiar</td>
<td>3 (25)</td>
</tr>
<tr>
<td>Somewhat familiar</td>
<td>7 (58)</td>
</tr>
<tr>
<td>Very familiar</td>
<td>2 (17)</td>
</tr>
<tr>
<td><strong>AI information source</strong></td>
<td></td>
</tr>
<tr>
<td>Family and friends</td>
<td>4 (33)</td>
</tr>
<tr>
<td>Career</td>
<td>5 (42)</td>
</tr>
<tr>
<td>Scholarly articles</td>
<td>2 (17)</td>
</tr>
<tr>
<td>Non–peer-reviewed articles</td>
<td>3 (25)</td>
</tr>
<tr>
<td>Social media</td>
<td>2 (17)</td>
</tr>
<tr>
<td>Other</td>
<td>3 (25)</td>
</tr>
<tr>
<td><strong>Medical care</strong></td>
<td></td>
</tr>
<tr>
<td><strong>Frequency of visiting an HCP\textsuperscript{b}</strong></td>
<td></td>
</tr>
<tr>
<td>Once a year</td>
<td>2 (17)</td>
</tr>
<tr>
<td>Fewer than 4 times a year</td>
<td>2 (17)</td>
</tr>
<tr>
<td>4 to 6 times a year</td>
<td>8 (66)</td>
</tr>
<tr>
<td><strong>Type of HCP</strong></td>
<td></td>
</tr>
<tr>
<td>Cardiologist</td>
<td>3 (25)</td>
</tr>
<tr>
<td>General practitioner</td>
<td>10 (83)</td>
</tr>
<tr>
<td>Ophthalmologist</td>
<td>3 (25)</td>
</tr>
<tr>
<td>Physiotherapist</td>
<td>3 (25)</td>
</tr>
<tr>
<td>Surgeon</td>
<td>3 (25)</td>
</tr>
<tr>
<td>Other</td>
<td>8 (67)</td>
</tr>
</tbody>
</table>

\textsuperscript{a}AI: artificial intelligence.

\textsuperscript{b}HCP: health care professional.
Table 2. Summary of key themes.

<table>
<thead>
<tr>
<th>Theme and quote</th>
<th>Significance</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Theme 1: cultivating patients' trust</strong></td>
<td>Transparent communication and acknowledging patient concerns and needs are imperative in fostering patients’ trust.</td>
</tr>
<tr>
<td><strong>Subtheme: providing safe and compassionate care</strong></td>
<td></td>
</tr>
<tr>
<td>“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]</td>
<td></td>
</tr>
<tr>
<td>“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]</td>
<td></td>
</tr>
<tr>
<td><strong>Subtheme: achieving transparency in care decisions</strong></td>
<td>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.</td>
</tr>
<tr>
<td>“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]</td>
<td></td>
</tr>
<tr>
<td><strong>Theme 2: fostering patient engagement</strong></td>
<td>Participants highlighted the importance of HCPs engaging them during the clinical decision-making process and providing an opportunity for them to share their thoughts and perspectives.</td>
</tr>
<tr>
<td><strong>Subtheme: enabling patients to be coleders in their care</strong></td>
<td></td>
</tr>
<tr>
<td>“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]</td>
<td></td>
</tr>
<tr>
<td>“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]</td>
<td></td>
</tr>
<tr>
<td>Theme and quote</td>
<td>Significance</td>
</tr>
<tr>
<td>----------------</td>
<td>--------------</td>
</tr>
<tr>
<td>• “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]</td>
<td>• 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.</td>
</tr>
<tr>
<td>• “…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]</td>
<td>• 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.</td>
</tr>
<tr>
<td>• “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]</td>
<td>• In addition, educating patients on the fundamentals of AI and other technologies can increase their confidence.</td>
</tr>
</tbody>
</table>

**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]

**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]

**Subtheme: ensuring AI technologies used in clinical contexts are equitable and inclusive**

- Quality assurance and validation of AI technologies are pivotal in ensuring the confidentiality of patient data and protecting them against nefarious acts.
<table>
<thead>
<tr>
<th>Theme and quote</th>
<th>Significance</th>
</tr>
</thead>
</table>
| - "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]|  
- 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. [ID 7]

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 semi-structured 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.

<table>
<thead>
<tr>
<th>Bloom taxonomy domain</th>
<th>Competencies identified in the scoping review (Charow et al [34])</th>
<th>Competencies highlighted by participants in this study and the scoping review (Charow et al [34])</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Cognitive</strong></td>
<td>• Fundamentals of AI&lt;sup&gt;a&lt;/sup&gt;</td>
<td>• Ethics and legal issues</td>
</tr>
<tr>
<td></td>
<td>• Implementation of AI</td>
<td>• Data governance</td>
</tr>
<tr>
<td></td>
<td>• Big data</td>
<td></td>
</tr>
<tr>
<td></td>
<td>• Data science, machine learning, and statistics</td>
<td></td>
</tr>
<tr>
<td></td>
<td>• Multidisciplinary collaboration</td>
<td></td>
</tr>
<tr>
<td></td>
<td>• Strengths and limitations of AI</td>
<td></td>
</tr>
<tr>
<td></td>
<td>• Predictive analytics</td>
<td></td>
</tr>
<tr>
<td></td>
<td>• Economic considerations</td>
<td></td>
</tr>
<tr>
<td></td>
<td>• EHR&lt;sup&gt;b&lt;/sup&gt; fundamentals</td>
<td></td>
</tr>
<tr>
<td><strong>Psychomotor</strong></td>
<td>• Analytical</td>
<td>• Interpretation</td>
</tr>
<tr>
<td></td>
<td>• Problem solving</td>
<td>• Communication</td>
</tr>
<tr>
<td></td>
<td>• Product development</td>
<td>• Critical appraisal</td>
</tr>
<tr>
<td></td>
<td>• Data visualization</td>
<td>• Medical decision-making</td>
</tr>
<tr>
<td><strong>Affective</strong></td>
<td>• Change management</td>
<td>• Cultivation of compassion and empathy</td>
</tr>
<tr>
<td></td>
<td>• Adoption of AI</td>
<td></td>
</tr>
<tr>
<td></td>
<td>• Perceptions of humanistic AI-enabled care</td>
<td></td>
</tr>
<tr>
<td></td>
<td>• Create and sustain a culture of trust and transparency with stakeholders and patients</td>
<td></td>
</tr>
</tbody>
</table>

<sup>a</sup>AI: artificial intelligence.  
<sup>b</sup>EHR: 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 des santé canadiennes 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.

---

https://ai.jmir.org/2023/1/e40973
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.

References


Abbreviations

AI: artificial intelligence
HCP: health care professional
Developing an Inpatient Electronic Medical Record Phenotype for Hospital-Acquired Pressure Injuries: Case Study Using Natural Language Processing Models

Elvira Nurmambetova1*, BHSc; Jie Pan1,2*, PhD; Zilong Zhang1*, MSc; Guosong Wu1,2, PhD; Seungwon Lee1,2,3, MPH, PhD; Danielle A Southern1,2, MSc; Elliot A Martin1,3, PhD; Chester Ho4, MD; Yuan Xu1,2,5,6, MD, PhD; Cathy A Eastwood1,2, 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

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 F1-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 F1-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).
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.
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].
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 F1-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. F1-score is a measure of accuracy through a combination of sensitivity and PPV. F1 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 F1-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).

<table>
<thead>
<tr>
<th>Characteristics</th>
<th>All</th>
<th>Patients with pressure injury (n=97)</th>
<th>Patients without pressure injury (n=183)</th>
<th>P value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Female, n (%)</td>
<td>127 (45.36)</td>
<td>42 (43.30)</td>
<td>85 (46.45)</td>
<td>.62</td>
</tr>
<tr>
<td>Age (years), median (IQR)</td>
<td>68 (55-79)</td>
<td>68 (59-80)</td>
<td>67 (53-79)</td>
<td>.10</td>
</tr>
<tr>
<td>Length of stay (days), median (IQR)</td>
<td>46 (22-104)</td>
<td>48 (28-96)</td>
<td>46 (19-109)</td>
<td>.37</td>
</tr>
<tr>
<td>Cerebrovascular disease, n (%)</td>
<td>137 (48.93)</td>
<td>50 (51.55)</td>
<td>87 (47.54)</td>
<td>.52</td>
</tr>
<tr>
<td>Chronic obstructive pulmonary disease, n (%)</td>
<td>54 (19.29)</td>
<td>26 (26.80)</td>
<td>28 (15.30)</td>
<td>.02</td>
</tr>
<tr>
<td>Congestive heart failure, n (%)</td>
<td>66 (23.57)</td>
<td>30 (30.93)</td>
<td>36 (19.67)</td>
<td>.04</td>
</tr>
<tr>
<td>Myocardial infarction, n (%)</td>
<td>33 (11.79)</td>
<td>9 (9.28)</td>
<td>24 (13.11)</td>
<td>.34</td>
</tr>
<tr>
<td>Dementia, n (%)</td>
<td>27 (9.64)</td>
<td>10 (10.31)</td>
<td>17 (9.29)</td>
<td>.79</td>
</tr>
<tr>
<td>Hemiplegia or paraplegia, n (%)</td>
<td>119 (42.50)</td>
<td>39 (40.21)</td>
<td>80 (43.72)</td>
<td>.57</td>
</tr>
<tr>
<td>Leukemia, n (%)</td>
<td>2 (0.71)</td>
<td>1 (1.03)</td>
<td>1 (0.55)</td>
<td>.62</td>
</tr>
<tr>
<td>Lymphoma, n (%)</td>
<td>4 (1.43)</td>
<td>2 (2.06)</td>
<td>2 (1.09)</td>
<td>.50</td>
</tr>
<tr>
<td>Peptic ulcer disease, n (%)</td>
<td>41 (14.64)</td>
<td>21 (21.65)</td>
<td>20 (10.93)</td>
<td>.02</td>
</tr>
<tr>
<td>Moderate or severe renal disease, n (%)</td>
<td>55 (19.64)</td>
<td>28 (28.87)</td>
<td>27 (14.75)</td>
<td>.005</td>
</tr>
<tr>
<td>Liver disease, n (%)</td>
<td>23 (8.21)</td>
<td>8 (8.25)</td>
<td>15 (8.20)</td>
<td>.99</td>
</tr>
<tr>
<td>Diabetes mellitus, n (%)</td>
<td>91 (32.50)</td>
<td>39 (40.21)</td>
<td>52 (28.42)</td>
<td>.045</td>
</tr>
<tr>
<td>Solid tumor, n (%)</td>
<td>50 (17.86)</td>
<td>19 (19.59)</td>
<td>31 (16.94)</td>
<td>.58</td>
</tr>
<tr>
<td>Connective tissue, n (%) disease</td>
<td>51 (18.21)</td>
<td>24 (24.74)</td>
<td>27 (14.75)</td>
<td>.04</td>
</tr>
<tr>
<td>History of smoking, n (%)</td>
<td>118 (42.14)</td>
<td>45 (46.39)</td>
<td>73 (39.89)</td>
<td>.30</td>
</tr>
<tr>
<td>Currently smoking, n (%)</td>
<td>37 (13.21)</td>
<td>11 (11.34)</td>
<td>26 (14.21)</td>
<td>.50</td>
</tr>
<tr>
<td>History of illicit drug use, n (%)</td>
<td>15 (5.36)</td>
<td>7 (7.22)</td>
<td>8 (4.37)</td>
<td>.32</td>
</tr>
<tr>
<td>Currently use illicit drugs, n (%)</td>
<td>8 (2.86)</td>
<td>3 (3.09)</td>
<td>5 (2.73)</td>
<td>.85</td>
</tr>
</tbody>
</table>

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.

<table>
<thead>
<tr>
<th>Document type</th>
<th>All (N=280)</th>
<th>Patients with HAPI (n=97)</th>
<th>Patients without HAPI (n=183)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Patient assessment</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Number of patients with this type of document, n (%)</td>
<td>142 (50.71)</td>
<td>60 (61.86)</td>
<td>82 (44.81)</td>
</tr>
<tr>
<td>Number of notes per patient, median (IQR)</td>
<td>1.00 (0.00-17.00)</td>
<td>12.00 (0.00-18.00)</td>
<td>0.00 (0.00-16.00)</td>
</tr>
<tr>
<td>Word count per note, median (IQR)</td>
<td>376.00 (306.00-430.00)</td>
<td>385.00 (311.00-441.00)</td>
<td>370.00 (303.00-421.00)</td>
</tr>
<tr>
<td><strong>Patient assessment neuro</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Number of patients with this type of document, n (%)</td>
<td>141 (50.36)</td>
<td>39 (40.21)</td>
<td>102 (55.74)</td>
</tr>
<tr>
<td>Number of notes per patient, median (IQR)</td>
<td>2.50 (0.00-13.00)</td>
<td>0.00 (0.00-13.00)</td>
<td>8.00 (0.00-13.00)</td>
</tr>
<tr>
<td>Word count per note, median (IQR)</td>
<td>428.00 (343.0-498.0)</td>
<td>424.00 (324.0-493.50)</td>
<td>430.00 (346.00-499.00)</td>
</tr>
<tr>
<td><strong>Patient care</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Number of patients with this type of document, n (%)</td>
<td>280 (100)</td>
<td>97 (100)</td>
<td>183(100)</td>
</tr>
<tr>
<td>Number of notes per patient, median (IQR)</td>
<td>16.00 (13.00-21.20)</td>
<td>16.00 (13.00-19.00)</td>
<td>16.00 (13.00-23.50)</td>
</tr>
<tr>
<td>Word count per note, median (IQR)</td>
<td>138.00 (72.00-179.00)</td>
<td>147.00 (91.00-185.00)</td>
<td>133.00 (66.00-176.00)</td>
</tr>
</tbody>
</table>

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 EMR 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\textsuperscript{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\textsuperscript{b} algorithms such that the model was trained to mimic its specificity.

<table>
<thead>
<tr>
<th>Model</th>
<th>Sensitivity % (95% CI)</th>
<th>Specificity % (95% CI)</th>
<th>PPV\textsuperscript{c} % (95% CI)</th>
<th>NPV\textsuperscript{d} % (95% CI)</th>
<th>F\textsubscript{1}-score % (95% CI)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Specificity near 0.988</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ICD (Ho et al [1])</td>
<td>0.277 (0.174-0.380)</td>
<td>0.988 (0.963-1.013)</td>
<td>0.917 (0.854-0.980)</td>
<td>0.739 (0.638-0.840)</td>
<td>0.425 (0.312-0.538)</td>
</tr>
<tr>
<td>TF-IDF+random forest\textsuperscript{f}</td>
<td>0.464 (0.365-0.563)</td>
<td>0.984 (0.965-1.000)</td>
<td>0.938 (0.869-1.000)</td>
<td>0.776 (0.722-0.830)</td>
<td>0.612 (0.473-0.751)</td>
</tr>
<tr>
<td>Count+random forest</td>
<td>0.412 (0.314-0.510)</td>
<td>0.978 (0.957-0.999)</td>
<td>0.909 (0.824-0.994)</td>
<td>0.758 (0.704-0.813)</td>
<td>0.550 (0.361-0.739)</td>
</tr>
<tr>
<td>TF-IDF+XGBoost\textsuperscript{g}</td>
<td>0.309 (0.217-0.401)</td>
<td>0.973 (0.949-0.996)</td>
<td>0.857 (0.741-0.973)</td>
<td>0.727 (0.671-0.782)</td>
<td>0.450 (0.340-0.559)</td>
</tr>
<tr>
<td>Word Embedding+BERT\textsuperscript{h}</td>
<td>0.268 (0.180-0.356)</td>
<td>0.978 (0.957-0.999)</td>
<td>0.867 (0.745-0.988)</td>
<td>0.716 (0.660-0.772)</td>
<td>0.394 (0.207-0.580)</td>
</tr>
<tr>
<td><strong>Specificity near 0.959</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ICD (Ho et al [1])</td>
<td>0.328 (0.220-0.436)</td>
<td>0.959 (0.913-1.000)</td>
<td>0.793 (0.700-0.886)</td>
<td>0.746 (0.646-0.846)</td>
<td>0.464 (0.350-0.578)</td>
</tr>
<tr>
<td>TF-IDF+random forest</td>
<td>0.546 (0.447-0.645)</td>
<td>0.951 (0.919-0.982)</td>
<td>0.855 (0.767-0.943)</td>
<td>0.798 (0.745-0.851)</td>
<td>0.665 (0.577-0.753)</td>
</tr>
<tr>
<td>Count+random forest</td>
<td>0.423 (0.324-0.521)</td>
<td>0.956 (0.927-0.986)</td>
<td>0.837 (0.733-0.940)</td>
<td>0.758 (0.702-0.813)</td>
<td>0.546 (0.359-0.733)</td>
</tr>
<tr>
<td>TF-IDF+XGBoost</td>
<td>0.423 (0.324-0.521)</td>
<td>0.956 (0.927-0.986)</td>
<td>0.837 (0.733-0.940)</td>
<td>0.758 (0.702-0.813)</td>
<td>0.552 (0.404-0.699)</td>
</tr>
<tr>
<td>Word embedding+BERT</td>
<td>0.289 (0.198-0.379)</td>
<td>0.967 (0.941-0.993)</td>
<td>0.824 (0.695-0.952)</td>
<td>0.720 (0.663-0.776)</td>
<td>0.420 (0.280-0.560)</td>
</tr>
</tbody>
</table>

\textsuperscript{a}NLP: natural language processing.
\textsuperscript{b}ICD: International Classification of Diseases.
\textsuperscript{c}PPV: positive predictive value.
\textsuperscript{d}NPV: negative predictive value.
\textsuperscript{e}TF-IDF: term frequency-inverse document frequency.
\textsuperscript{f}The best model.
\textsuperscript{g}XGBoost: extreme gradient boosting.
\textsuperscript{h}BERT: 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 F1-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_v21e41264_app1.docx]

Multimedia Appendix 2
Linguistic inquiry and word count analysis.
[DOCX File, 16 KB - ai_v21e41264_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_v21e41264_app3.docx]

References


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
Deep Learning to Detect Pancreatic Cystic Lesions on Abdominal Computed Tomography Scans: Development and Validation Study

Maria Montserrat Duh†, MD; Neus Torra-Ferrer†, MD; Merixell Riera-Marín‡, MSc; Dídac Cumelles‡, BSc; Júlia Rodríguez-Comas‡, PhD; Javier García López‡, PhD; Mª Teresa Fernández Planas†*, MD

†Department of Radiology, Consorci Sanitari del Maresme (Hospital de Mataró), Mataró, Spain
‡Scientific and Technical Department, Sycai Technologies SL, Barcelona, Spain
*these authors contributed equally

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.

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].

https://ai.jmir.org/2023/1/e40702

(JMIR AI 2023;2:e40702) doi:10.2196/40702

KEYWORDS
deep learning; pancreatic cystic lesion; neural networks; precursor lesions; pancreatic cancer; computed tomography; magnetic resonance; cancer; radiologist; technology
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 prothesis 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.

<table>
<thead>
<tr>
<th>Diagnosis</th>
<th>Values, n (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Serous cystadenoma</td>
<td>42 (12.5)</td>
</tr>
<tr>
<td>Intraductal papillary mucinous neoplasm</td>
<td>154 (46)</td>
</tr>
<tr>
<td>Mucinous cystic neoplasms</td>
<td>5 (1.5)</td>
</tr>
<tr>
<td>Pseudocyst</td>
<td>82 (24.5)</td>
</tr>
<tr>
<td>Cyst</td>
<td>14 (4.2)</td>
</tr>
<tr>
<td>No cyst</td>
<td>38 (11.3)</td>
</tr>
</tbody>
</table>

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 helixes 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 helixes 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 ($\alpha$) and the input feature maps came from the previous stage of the decoding part ($\chi$; 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 $F$, where $F'$ 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 $g$, 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.
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 $512 \times 512$ to $240 \times 240$ 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 $160 \times 160$ of this central crop, and therefore, the inference consisted of iterating around this central crop of multiple inferences of patches of $160 \times 160$.

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.

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.

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.

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.
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 determination 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 Sycai Technologies and declare significant ownership. MRM and DC are employed by Sycai Technologies. The other authors report no conflicts of interest.

References
21. Attribution 4.0 International (CC BY 4.0). Creative Commons. URL: https://creativecommons.org/licenses/by/4.0/ [accessed [2023-03-16]]


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
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
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 F₁-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

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.

<table>
<thead>
<tr>
<th></th>
<th>Depressed users</th>
<th>Control users</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total users</td>
<td>70</td>
<td>700</td>
</tr>
<tr>
<td>Total posts</td>
<td>36,826</td>
<td>364,747</td>
</tr>
<tr>
<td>Total words</td>
<td>1,742,388</td>
<td>8,188,090</td>
</tr>
<tr>
<td>Average posts per user</td>
<td>526.1</td>
<td>521.1</td>
</tr>
<tr>
<td>Average words per post</td>
<td>47.3</td>
<td>22.4</td>
</tr>
<tr>
<td>Shortest post (words)</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Longest post (words)</td>
<td>2642</td>
<td>1894</td>
</tr>
</tbody>
</table>

**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.

<table>
<thead>
<tr>
<th>User, word</th>
<th>Score</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Depressed users</strong></td>
<td></td>
</tr>
<tr>
<td>people</td>
<td>338,131.45</td>
</tr>
<tr>
<td>know</td>
<td>164,368.51</td>
</tr>
<tr>
<td>thing</td>
<td>150,440.49</td>
</tr>
<tr>
<td>feel</td>
<td>118,483.23</td>
</tr>
<tr>
<td>time</td>
<td>97,250.09</td>
</tr>
<tr>
<td>woman</td>
<td>96,165.35</td>
</tr>
<tr>
<td>go</td>
<td>79,611.79</td>
</tr>
<tr>
<td>want</td>
<td>75,379.17</td>
</tr>
<tr>
<td>life</td>
<td>67,769.01</td>
</tr>
<tr>
<td>relationship</td>
<td>62,606.64</td>
</tr>
<tr>
<td><strong>Control users</strong></td>
<td></td>
</tr>
<tr>
<td>game</td>
<td>42,234.94</td>
</tr>
<tr>
<td>trade</td>
<td>39,445.65</td>
</tr>
<tr>
<td>key</td>
<td>30,031.17</td>
</tr>
<tr>
<td>team</td>
<td>24,333.73</td>
</tr>
<tr>
<td>play</td>
<td>17,389.38</td>
</tr>
<tr>
<td>player</td>
<td>16,186.61</td>
</tr>
<tr>
<td>shiny</td>
<td>14,032.27</td>
</tr>
<tr>
<td>hatch</td>
<td>13,265.87</td>
</tr>
<tr>
<td>thank</td>
<td>10,177.49</td>
</tr>
<tr>
<td>add</td>
<td>10,005.14</td>
</tr>
</tbody>
</table>

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.**

<table>
<thead>
<tr>
<th>Variation</th>
<th>Depressed users</th>
<th>Matching control users per depressed user</th>
<th>Total users</th>
</tr>
</thead>
<tbody>
<tr>
<td>Variation 1</td>
<td>70</td>
<td>1</td>
<td>140</td>
</tr>
<tr>
<td>Variation 2</td>
<td>70</td>
<td>3</td>
<td>280</td>
</tr>
<tr>
<td>Variation 3</td>
<td>70</td>
<td>5</td>
<td>420</td>
</tr>
<tr>
<td>Variation 4</td>
<td>70</td>
<td>10</td>
<td>770</td>
</tr>
</tbody>
</table>

**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.

<table>
<thead>
<tr>
<th></th>
<th>Depressed users</th>
<th>Matching control users per depressed user</th>
<th>Total users</th>
<th>Weeks of posts included before estimated diagnosis date</th>
</tr>
</thead>
<tbody>
<tr>
<td>Variation 1</td>
<td>56</td>
<td>1</td>
<td>112</td>
<td>4, 8, 12, 16, 20, and 24</td>
</tr>
<tr>
<td>Variation 2</td>
<td>56</td>
<td>3</td>
<td>224</td>
<td>4, 8, 12, 16, 20, and 24</td>
</tr>
</tbody>
</table>
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 naïve 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.

<table>
<thead>
<tr>
<th>Language Model</th>
<th>Positive predict. value, mean (SD)</th>
<th>Sensitivity, mean (SD)</th>
<th>F1-score, mean (SD)</th>
</tr>
</thead>
<tbody>
<tr>
<td>SVM using TF-IDF</td>
<td>0.637 (N/A)</td>
<td>0.557 (N/A)</td>
<td>0.590 (N/A)</td>
</tr>
<tr>
<td>SVM using word embeddings</td>
<td>0.558 (N/A)</td>
<td>0.543 (N/A)</td>
<td>0.548 (N/A)</td>
</tr>
<tr>
<td>SVM using TF-IDF and word embeddings</td>
<td>0.673 (N/A)</td>
<td>0.557 (N/A)</td>
<td>0.596 (N/A)</td>
</tr>
<tr>
<td>BERT LM</td>
<td>0.638 (0.021)</td>
<td>0.805 (0.022)</td>
<td>0.709 (0.012)</td>
</tr>
<tr>
<td>ALBERT LM</td>
<td>0.606 (0.008)</td>
<td>0.786 (0.015)</td>
<td>0.683 (0.010)</td>
</tr>
<tr>
<td>BioBERT LM</td>
<td>0.601 (0.005)</td>
<td>0.862 (0.022)</td>
<td>0.707 (0.005)</td>
</tr>
<tr>
<td>Longformer LM</td>
<td>0.633 (0.009)</td>
<td>0.838 (0.036)</td>
<td>0.719 (0.018)</td>
</tr>
<tr>
<td>MentalBERT LM</td>
<td>0.660 (0.019)</td>
<td>0.848 (0.008)</td>
<td>0.738 (0.013)</td>
</tr>
<tr>
<td>MentalRoBERTa LM</td>
<td>0.629 (0.002)</td>
<td>0.819 (0.022)</td>
<td>0.709 (0.006)</td>
</tr>
<tr>
<td>Naive baseline—all depression</td>
<td>0.500 (N/A)</td>
<td>1.000 (N/A)</td>
<td>0.667 (N/A)</td>
</tr>
</tbody>
</table>

*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>
<thead>
<tr>
<th>Model Type</th>
<th>Positive predictive value, mean (SD)</th>
<th>Sensitivity, mean (SD)</th>
<th>$F_1$-score, mean (SD)</th>
</tr>
</thead>
<tbody>
<tr>
<td>SVM(^b) using TF-IDF(^c)</td>
<td>0.800 (N/A(^d))</td>
<td>0.086 (N/A)</td>
<td>0.153 (N/A)</td>
</tr>
<tr>
<td>SVM using word embeddings</td>
<td>0.411 (N/A)</td>
<td>0.529 (N/A)</td>
<td>0.459 (N/A)</td>
</tr>
<tr>
<td>SVM using TF-IDF and word embeddings</td>
<td>0.800 (N/A)</td>
<td>0.057 (N/A)</td>
<td>0.107 (N/A)</td>
</tr>
<tr>
<td>BERT(^e) LM(^f)</td>
<td>0.653 (0.033)</td>
<td>0.481 (0.022)</td>
<td>0.546 (0.025)</td>
</tr>
<tr>
<td>ALBERT(^g) LM</td>
<td>0.652 (0.034)</td>
<td>0.476 (0.009)</td>
<td>0.547 (0.018)</td>
</tr>
<tr>
<td>BioBERT LM</td>
<td>0.654 (0.028)</td>
<td>0.410 (0.030)</td>
<td>0.496 (0.020)</td>
</tr>
<tr>
<td>Longformer LM</td>
<td>0.653 (0.036)</td>
<td>0.476 (0.036)</td>
<td>0.534 (0.031)</td>
</tr>
<tr>
<td>MentalBERT LM</td>
<td>0.657 (0.034)</td>
<td>0.509 (0.008)</td>
<td>0.562 (0.016)</td>
</tr>
<tr>
<td>MentalRoBERTa LM</td>
<td>0.614 (0.023)</td>
<td>0.471 (0.015)</td>
<td>0.522 (0.002)</td>
</tr>
<tr>
<td>Naive baseline—all depression</td>
<td>0.250 (N/A)</td>
<td>1.000 (N/A)</td>
<td>0.167 (N/A)</td>
</tr>
</tbody>
</table>

\(^a\)Language model experiments were run 3 times each, therefore both mean and SD scores are provided.
\(^b\)SVM: support vector machine.
\(^c\)TF-IDF: term frequency–inverse document frequency.
\(^d\)N/A: not applicable.
\(^e\)BERT: Bidirectional Encoder Representations from Transformers.
\(^f\)LM: language model.
\(^g\)ALBERT: A Lite Bidirectional Encoder Representations from Transformers.

<table>
<thead>
<tr>
<th>Model Type</th>
<th>Positive predictive value, mean (SD)</th>
<th>Sensitivity, mean (SD)</th>
<th>$F_1$-score, mean (SD)</th>
</tr>
</thead>
<tbody>
<tr>
<td>SVM(^b) using TF-IDF(^c)</td>
<td>0.400 (N/A(^d))</td>
<td>0.029 (N/A)</td>
<td>0.053 (N/A)</td>
</tr>
<tr>
<td>SVM using word embeddings</td>
<td>0.309 (N/A)</td>
<td>0.471 (N/A)</td>
<td>0.372 (N/A)</td>
</tr>
<tr>
<td>SVM using TF-IDF and word embeddings</td>
<td>0.200 (N/A)</td>
<td>0.014 (N/A)</td>
<td>0.027 (N/A)</td>
</tr>
<tr>
<td>BERT(^e) LM(^f)</td>
<td>0.615 (0.028)</td>
<td>0.290 (0.022)</td>
<td>0.379 (0.017)</td>
</tr>
<tr>
<td>ALBERT(^g) LM</td>
<td>0.555 (0.030)</td>
<td>0.281 (0.009)</td>
<td>0.354 (0.006)</td>
</tr>
<tr>
<td>BioBERT LM</td>
<td>0.627 (0.034)</td>
<td>0.252 (0.021)</td>
<td>0.331 (0.027)</td>
</tr>
<tr>
<td>Longformer LM</td>
<td>0.624 (0.108)</td>
<td>0.286 (0.038)</td>
<td>0.363 (0.059)</td>
</tr>
<tr>
<td>MentalBERT LM</td>
<td>0.572 (0.002)</td>
<td>0.329 (0.043)</td>
<td>0.400 (0.040)</td>
</tr>
<tr>
<td>MentalRoBERTa LM</td>
<td>0.562 (0.027)</td>
<td>0.343 (0.000)</td>
<td>0.419 (0.010)</td>
</tr>
<tr>
<td>Naive baseline—all depression</td>
<td>0.167 (N/A)</td>
<td>1.000 (N/A)</td>
<td>0.286 (N/A)</td>
</tr>
</tbody>
</table>

\(^a\)Language model experiments were run 3 times each, therefore both mean and SD scores are provided.
\(^b\)SVM: support vector machine.
\(^c\)TF-IDF: term frequency–inverse document frequency.
\(^d\)N/A: not applicable.
\(^e\)BERT: Bidirectional Encoder Representations from Transformers.
\(^f\)LM: language model.
\(^g\)ALBERT: 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\(^a\).

<table>
<thead>
<tr>
<th>Model Description</th>
<th>Positive predictive value, mean (SD)</th>
<th>Sensitivity, mean (SD)</th>
<th>(F_1)-score, mean (SD)</th>
</tr>
</thead>
<tbody>
<tr>
<td>SVM(^b) using TF-IDF(^c)</td>
<td>0.000 (N/A(^d))</td>
<td>0.000 (N/A)</td>
<td>0.000 (N/A)</td>
</tr>
<tr>
<td>SVM using word embeddings</td>
<td>0.212 (N/A)</td>
<td>0.371 (N/A)</td>
<td>0.268 (N/A)</td>
</tr>
<tr>
<td>SVM using TF-IDF and word embeddings</td>
<td>0.000 (N/A)</td>
<td>0.000 (N/A)</td>
<td>0.000 (N/A)</td>
</tr>
<tr>
<td>BERT(^e) LM(^f)</td>
<td>0.100 (0.000)</td>
<td>0.014 (0.000)</td>
<td>0.025 (0.00)</td>
</tr>
<tr>
<td>ALBERT(^g) LM</td>
<td>0.089 (0.019)</td>
<td>0.014 (0.000)</td>
<td>0.025 (0.001)</td>
</tr>
<tr>
<td>BioBERT LM</td>
<td>0.067 (0.115)</td>
<td>0.005 (0.008)</td>
<td>0.009 (0.016)</td>
</tr>
<tr>
<td>Longformer LM</td>
<td>0.024 (0.019)</td>
<td>0.019 (0.033)</td>
<td>0.021 (0.037)</td>
</tr>
<tr>
<td>MentalBERT LM</td>
<td>0.167 (0.058)</td>
<td>0.014 (0.000)</td>
<td>0.026 (0.001)</td>
</tr>
<tr>
<td>MentalRoBERTa LM</td>
<td>0.272 (0.185)</td>
<td>0.034 (0.008)</td>
<td>0.057 (0.018)</td>
</tr>
<tr>
<td>Naive baseline—all depression</td>
<td>0.091 (N/A)</td>
<td>1.000 (N/A)</td>
<td>0.167 (N/A)</td>
</tr>
</tbody>
</table>

\(^a\)Language model experiments were run 3 times each, therefore both mean and SD scores are provided.

\(^b\)SVM: support vector machine.

\(^c\)TF-IDF: term frequency–inverse document frequency.

\(^d\)N/A: not applicable.

\(^e\)BERT: Bidirectional Encoder Representations from Transformers.

\(^f\)LM: language model.

\(^g\)ALBERT: A Lite Bidirectional Encoder Representations from Transformers.
Table 9. Depressed users most strongly misclassified in each variation of the preemptive depression identification experiment.

<table>
<thead>
<tr>
<th>Classifier</th>
<th>One depression user per control user (1:1)</th>
<th>One depression user per 3 control users (1:3)</th>
<th>One depression user per 5 control users (1:5)</th>
<th>One depression user per 10 control users (1:10)</th>
</tr>
</thead>
<tbody>
<tr>
<td>User</td>
<td>SVM using word embeddings</td>
<td>MentalRoBERTa LM</td>
<td>MentalBERT LM</td>
<td>MentalBERT LM</td>
</tr>
<tr>
<td>Control probability</td>
<td>0.93</td>
<td>0.94</td>
<td>0.99</td>
<td>0.98</td>
</tr>
<tr>
<td>Sum of post lengths in words</td>
<td>1696</td>
<td>1888</td>
<td>1696</td>
<td>55,897</td>
</tr>
<tr>
<td>Topic</td>
<td>news</td>
<td>sir-geo</td>
<td>news</td>
<td>good</td>
</tr>
<tr>
<td></td>
<td>hawaii</td>
<td>welcomed</td>
<td>hawaii</td>
<td>time</td>
</tr>
<tr>
<td></td>
<td>time</td>
<td>invite</td>
<td>time</td>
<td>people</td>
</tr>
<tr>
<td></td>
<td>dead</td>
<td>leave</td>
<td>dead</td>
<td>years</td>
</tr>
<tr>
<td></td>
<td>blue</td>
<td>warlock</td>
<td>blue</td>
<td>problem</td>
</tr>
<tr>
<td>Chief TF-IDF(^d) features</td>
<td>love</td>
<td>sir</td>
<td>love</td>
<td>good</td>
</tr>
<tr>
<td></td>
<td>minnesota</td>
<td>geo</td>
<td>minnesota</td>
<td>know</td>
</tr>
<tr>
<td></td>
<td>diablo</td>
<td>welcome</td>
<td>diablo</td>
<td>use</td>
</tr>
<tr>
<td></td>
<td>time</td>
<td>invite</td>
<td>time</td>
<td>make</td>
</tr>
<tr>
<td></td>
<td>man</td>
<td>warlock</td>
<td>man</td>
<td>make</td>
</tr>
<tr>
<td></td>
<td>bud</td>
<td>leave</td>
<td>bud</td>
<td>time</td>
</tr>
<tr>
<td></td>
<td>zoidberg</td>
<td>titan</td>
<td>zoidberg</td>
<td>thank</td>
</tr>
<tr>
<td></td>
<td>like</td>
<td>psn</td>
<td>like</td>
<td>link</td>
</tr>
<tr>
<td></td>
<td>month</td>
<td>run</td>
<td>month</td>
<td>want</td>
</tr>
<tr>
<td></td>
<td>hawaii</td>
<td>need</td>
<td>hawaii</td>
<td>try</td>
</tr>
</tbody>
</table>

Depressed vocabulary counts

<table>
<thead>
<tr>
<th></th>
<th>1</th>
<th>1</th>
<th>1</th>
<th>64</th>
</tr>
</thead>
<tbody>
<tr>
<td>people</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>64</td>
</tr>
<tr>
<td>know</td>
<td>6</td>
<td>0</td>
<td>6</td>
<td>93</td>
</tr>
<tr>
<td>thing</td>
<td>3</td>
<td>0</td>
<td>3</td>
<td>35</td>
</tr>
<tr>
<td>feel</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>10</td>
</tr>
<tr>
<td>time</td>
<td>5</td>
<td>8</td>
<td>5</td>
<td>99</td>
</tr>
<tr>
<td>woman</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>7</td>
</tr>
<tr>
<td>go</td>
<td>3</td>
<td>0</td>
<td>3</td>
<td>54</td>
</tr>
<tr>
<td>want</td>
<td>3</td>
<td>1</td>
<td>3</td>
<td>71</td>
</tr>
<tr>
<td>life</td>
<td>2</td>
<td>0</td>
<td>2</td>
<td>28</td>
</tr>
<tr>
<td>relationship</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>2</td>
</tr>
</tbody>
</table>

Control vocabulary counts

<table>
<thead>
<tr>
<th></th>
<th>0</th>
<th>1</th>
<th>0</th>
<th>9</th>
</tr>
</thead>
<tbody>
<tr>
<td>game</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>9</td>
</tr>
<tr>
<td>trade</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>2</td>
</tr>
<tr>
<td>key</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>4</td>
</tr>
<tr>
<td>team</td>
<td>2</td>
<td>3</td>
<td>2</td>
<td>4</td>
</tr>
<tr>
<td>play</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>35</td>
</tr>
<tr>
<td>player</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>8</td>
</tr>
<tr>
<td>shiny</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>hatch</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>thank</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>15</td>
</tr>
<tr>
<td>add</td>
<td>0</td>
<td>2</td>
<td>0</td>
<td>14</td>
</tr>
</tbody>
</table>

\(^a\)Lexical properties of those users’ posts are provided.

\(^b\)LM: language model.

\(^c\)SVM: support vector machine.

\(^d\)TF-IDF: term frequency–inverse document frequency.
Table 10. Control users most strongly misclassified in each variation of the preemptive depression identification experiment.\(^{a}\)

<table>
<thead>
<tr>
<th>Classifier</th>
<th>One depression user per control user (1:1)</th>
<th>One depression user per 3 control users (1:3)</th>
<th>One depression user per 5 control users (1:5)</th>
<th>One depression user per 10 control users (1:10)</th>
</tr>
</thead>
<tbody>
<tr>
<td>User</td>
<td>MentalBERT LMc</td>
<td>MentalBERT LM</td>
<td>MentalRoBERTa LM</td>
<td>SVMc using Word embeddings</td>
</tr>
<tr>
<td>Depressed probability</td>
<td>0.99</td>
<td>0.95</td>
<td>0.91</td>
<td>0.91</td>
</tr>
<tr>
<td>Sum of post lengths in words</td>
<td>1513</td>
<td>8489</td>
<td>8489</td>
<td>1595</td>
</tr>
<tr>
<td>Topic</td>
<td>elo</td>
<td>god</td>
<td>god</td>
<td>people</td>
</tr>
<tr>
<td></td>
<td>play</td>
<td>jesus</td>
<td>jesus</td>
<td>shit</td>
</tr>
<tr>
<td></td>
<td>team</td>
<td>people</td>
<td>people</td>
<td>reddit</td>
</tr>
<tr>
<td></td>
<td>bronze</td>
<td>good</td>
<td>good</td>
<td>guy</td>
</tr>
<tr>
<td></td>
<td>games</td>
<td>life</td>
<td>life</td>
<td>man</td>
</tr>
<tr>
<td>Chief TF-IDF(^{d}) features</td>
<td>team</td>
<td>god</td>
<td>god</td>
<td>say</td>
</tr>
<tr>
<td></td>
<td>just</td>
<td>think</td>
<td>think</td>
<td>thank</td>
</tr>
<tr>
<td></td>
<td>suck</td>
<td>way</td>
<td>way</td>
<td>say</td>
</tr>
<tr>
<td></td>
<td>elo</td>
<td>thing</td>
<td>thing</td>
<td>way</td>
</tr>
<tr>
<td></td>
<td>play</td>
<td>try</td>
<td>try</td>
<td>guy</td>
</tr>
<tr>
<td></td>
<td>game</td>
<td>know</td>
<td>know</td>
<td>people</td>
</tr>
<tr>
<td></td>
<td>like</td>
<td>jesus</td>
<td>jesus</td>
<td>reddit</td>
</tr>
<tr>
<td></td>
<td>good</td>
<td>people</td>
<td>people</td>
<td>man</td>
</tr>
<tr>
<td></td>
<td>sydiko</td>
<td>say</td>
<td>say</td>
<td>tell</td>
</tr>
<tr>
<td></td>
<td>win</td>
<td>like</td>
<td>like</td>
<td>watch</td>
</tr>
</tbody>
</table>

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

\(^{a}\)Lexical properties of those users’ posts are provided.
\(^{b}\)LM: language model.
\(^{c}\)SVM: Support Vector Machine.
\(^{d}\)TF-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\(^a\).

<table>
<thead>
<tr>
<th>Time period</th>
<th>Positive predictive value, mean (SD)</th>
<th>Sensitivity, mean (SD)</th>
<th>(F_1)-score, mean (SD)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Last 4 weeks</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>BERT(^b) LM(^c)</td>
<td>0.575 (0.027)</td>
<td>0.830 (0.039)</td>
<td>0.675 (0.023)</td>
</tr>
<tr>
<td>MentalBERT LM</td>
<td>0.612 (0.026)</td>
<td>0.835 (0.026)</td>
<td>0.698 (0.017)</td>
</tr>
<tr>
<td><strong>Last 8 weeks</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>BERT LM</td>
<td>0.598 (0.026)</td>
<td>0.854 (0.071)</td>
<td>0.700 (0.037)</td>
</tr>
<tr>
<td>MentalBERT LM</td>
<td>0.603 (0.020)</td>
<td>0.842 (0.047)</td>
<td>0.699 (0.022)</td>
</tr>
<tr>
<td><strong>Last 12 weeks</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>BERT LM</td>
<td>0.605 (0.014)</td>
<td>0.912 (0.018)</td>
<td>0.726 (0.015)</td>
</tr>
<tr>
<td>MentalBERT LM</td>
<td>0.600 (0.013)</td>
<td>0.888 (0.010)</td>
<td>0.715 (0.008)</td>
</tr>
<tr>
<td><strong>Last 16 weeks</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>BERT LM</td>
<td>0.570 (0.009)</td>
<td>0.865 (0.026)</td>
<td>0.684 (0.007)</td>
</tr>
<tr>
<td>MentalBERT LM</td>
<td>0.575 (0.009)</td>
<td>0.907 (0.028)</td>
<td>0.703 (0.016)</td>
</tr>
<tr>
<td><strong>Last 20 weeks</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>BERT LM</td>
<td>0.569 (0.023)</td>
<td>0.893 (0.036)</td>
<td>0.694 (0.025)</td>
</tr>
<tr>
<td>MentalBERT LM</td>
<td>0.578 (0.018)</td>
<td>0.882 (0.027)</td>
<td>0.696 (0.014)</td>
</tr>
<tr>
<td><strong>Last 24 weeks</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>BERT LM</td>
<td>0.565 (0.021)</td>
<td>0.871 (0.027)</td>
<td>0.683 (0.010)</td>
</tr>
<tr>
<td>MentalBERT LM</td>
<td>0.591 (0.014)</td>
<td>0.890 (0.010)</td>
<td>0.707 (0.011)</td>
</tr>
<tr>
<td>All posts</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>BERT LM</td>
<td>0.627 (0.018)</td>
<td>0.824 (0.032)</td>
<td>0.710 (0.019)</td>
</tr>
<tr>
<td>MentalBERT LM</td>
<td>0.638 (0.009)</td>
<td>0.861 (0.000)</td>
<td>0.732 (0.006)</td>
</tr>
<tr>
<td>Naive baseline</td>
<td>0.500 (N/A(^d))</td>
<td>1.000 (N/A)</td>
<td>0.667 (N/A)</td>
</tr>
</tbody>
</table>

\(^a\)The 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.
\(^b\)BERT: Bidirectional Encoder Representations From Transformers.
\(^c\)LM: language model.
\(^d\)N/A: not applicable.
Table 12. Binary classification scores using 56 depressed users and 3 of their matched control users and 6 temporal post subsets\(^a\).

<table>
<thead>
<tr>
<th>Time Period</th>
<th>Positive predictive value, mean (SD)</th>
<th>Sensitivity, mean (SD)</th>
<th>(F_1)-score, mean (SD)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Last 4 weeks</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>BERT(^b) LM(^c)</td>
<td>0.480 (0.027)</td>
<td>0.538 (0.019)</td>
<td>0.489 (0.010)</td>
</tr>
<tr>
<td>MentalBERT LM</td>
<td>0.494 (0.019)</td>
<td>0.577 (0.009)</td>
<td>0.525 (0.007)</td>
</tr>
<tr>
<td><strong>Last 8 weeks</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>BERT LM</td>
<td>0.446 (0.032)</td>
<td>0.538 (0.036)</td>
<td>0.472 (0.035)</td>
</tr>
<tr>
<td>MentalBERT LM</td>
<td>0.427 (0.027)</td>
<td>0.524 (0.029)</td>
<td>0.461 (0.023)</td>
</tr>
<tr>
<td><strong>Last 12 weeks</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>BERT LM</td>
<td>0.498 (0.031)</td>
<td>0.619 (0.037)</td>
<td>0.543 (0.035)</td>
</tr>
<tr>
<td>MentalBERT LM</td>
<td>0.448 (0.007)</td>
<td>0.569 (0.017)</td>
<td>0.494 (0.009)</td>
</tr>
<tr>
<td><strong>Last 16 weeks</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>BERT LM</td>
<td>0.471 (0.010)</td>
<td>0.565 (0.021)</td>
<td>0.504 (0.011)</td>
</tr>
<tr>
<td>MentalBERT LM</td>
<td>0.481 (0.023)</td>
<td>0.643 (0.037)</td>
<td>0.541 (0.028)</td>
</tr>
<tr>
<td><strong>Last 20 weeks</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>BERT LM</td>
<td>0.475 (0.039)</td>
<td>0.577 (0.037)</td>
<td>0.510 (0.034)</td>
</tr>
<tr>
<td>MentalBERT LM</td>
<td>0.487 (0.018)</td>
<td>0.595 (0.011)</td>
<td>0.524 (0.009)</td>
</tr>
<tr>
<td><strong>Last 24 weeks</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>BERT LM</td>
<td>0.470 (0.033)</td>
<td>0.591 (0.036)</td>
<td>0.518 (0.033)</td>
</tr>
<tr>
<td>MentalBERT LM</td>
<td>0.501 (0.022)</td>
<td>0.591 (0.018)</td>
<td>0.536 (0.022)</td>
</tr>
<tr>
<td><strong>All posts</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>BERT LM</td>
<td>0.625 (0.021)</td>
<td>0.519 (0.032)</td>
<td>0.562 (0.015)</td>
</tr>
<tr>
<td>MentalBERT LM</td>
<td>0.588 (0.005)</td>
<td>0.508 (0.010)</td>
<td>0.540 (0.003)</td>
</tr>
<tr>
<td>Naive baseline</td>
<td>0.250 (N/A(^d))</td>
<td>1.000 (N/A)</td>
<td>0.400 (N/A)</td>
</tr>
</tbody>
</table>

\(^a\)The 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.

\(^b\)BERT: Bidirectional Encoder Representations From Transformer.

\(^c\)LM: language model.

\(^d\)N/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.

<table>
<thead>
<tr>
<th>Time span</th>
<th>Last 12 weeks</th>
<th>Last 12 weeks</th>
</tr>
</thead>
<tbody>
<tr>
<td>Classifier</td>
<td>BERT&lt;sup&gt;a&lt;/sup&gt; LM&lt;sup&gt;b&lt;/sup&gt;</td>
<td>BERT LM</td>
</tr>
<tr>
<td>User</td>
<td>d52</td>
<td>d52</td>
</tr>
<tr>
<td>Control probability</td>
<td>0.869</td>
<td>0.935</td>
</tr>
<tr>
<td>Sum of post lengths in words</td>
<td>1225</td>
<td>1225</td>
</tr>
<tr>
<td>Topic</td>
<td>england belgium hamster time team</td>
<td>england belgium hamster time team</td>
</tr>
<tr>
<td>Chief TF-IDF&lt;sup&gt;c&lt;/sup&gt; features</td>
<td>• thank</td>
<td>• thank</td>
</tr>
<tr>
<td></td>
<td>• team</td>
<td>• team</td>
</tr>
<tr>
<td></td>
<td>• player</td>
<td>• player</td>
</tr>
<tr>
<td></td>
<td>• help</td>
<td>• help</td>
</tr>
<tr>
<td></td>
<td>• time</td>
<td>• time</td>
</tr>
<tr>
<td></td>
<td>• goal</td>
<td>• goal</td>
</tr>
<tr>
<td></td>
<td>• cage</td>
<td>• cage</td>
</tr>
<tr>
<td></td>
<td>• post</td>
<td>• post</td>
</tr>
<tr>
<td></td>
<td>• second</td>
<td>• second</td>
</tr>
<tr>
<td></td>
<td>• start</td>
<td>• start</td>
</tr>
</tbody>
</table>

#### Depressed vocabulary counts

<table>
<thead>
<tr>
<th></th>
<th>People</th>
<th>Know</th>
<th>Thing</th>
<th>Feel</th>
<th>Time</th>
<th>Woman</th>
<th>Go</th>
<th>Want</th>
<th>Life</th>
<th>Relationship</th>
</tr>
</thead>
<tbody>
<tr>
<td>people</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>know</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>thing</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>feel</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>time</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>4</td>
</tr>
<tr>
<td>woman</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>go</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>want</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>life</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>relationship</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

#### Control vocabulary counts

<table>
<thead>
<tr>
<th></th>
<th>Game</th>
<th>Trade</th>
<th>Key</th>
<th>Team</th>
<th>Play</th>
<th>Player</th>
<th>Shiny</th>
<th>Hatch</th>
<th>Thank</th>
<th>Add</th>
</tr>
</thead>
<tbody>
<tr>
<td>game</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>trade</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>key</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>team</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>4</td>
</tr>
<tr>
<td>play</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>player</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>shiny</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>hatch</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>thank</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>add</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
</tbody>
</table>

<sup>a</sup>BERT: Bidirectional Encoder Representations From Transformers.
<sup>b</sup>LM: language model.
<sup>c</sup>TF-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.

<table>
<thead>
<tr>
<th>Time span</th>
<th>Last 12 weeks</th>
<th>Last 12 weeks per 3 control users (1:3)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Classifier</td>
<td>BERT(^a) LM(^b)</td>
<td>BERT LM</td>
</tr>
<tr>
<td>User</td>
<td>c481</td>
<td>c13</td>
</tr>
<tr>
<td>Depressed probability</td>
<td>0.963</td>
<td>0.917</td>
</tr>
<tr>
<td>Total length of posts in words</td>
<td>258</td>
<td>8489</td>
</tr>
<tr>
<td>Topic</td>
<td>● food</td>
<td>● god</td>
</tr>
<tr>
<td></td>
<td>● clove</td>
<td>● jesus</td>
</tr>
<tr>
<td></td>
<td>● tomorrow</td>
<td>● people</td>
</tr>
<tr>
<td></td>
<td>● pain</td>
<td>● good</td>
</tr>
<tr>
<td></td>
<td>● suspect</td>
<td>● life</td>
</tr>
<tr>
<td>Chief TF-IDF(^c) features</td>
<td>● reply</td>
<td>● god</td>
</tr>
<tr>
<td></td>
<td>● eat</td>
<td>● think</td>
</tr>
<tr>
<td></td>
<td>● food</td>
<td>● way</td>
</tr>
<tr>
<td></td>
<td>● cat</td>
<td>● thing</td>
</tr>
<tr>
<td></td>
<td>● clove</td>
<td>● try</td>
</tr>
<tr>
<td></td>
<td>● pain</td>
<td>● know</td>
</tr>
<tr>
<td></td>
<td>● suspect</td>
<td>● jesus</td>
</tr>
<tr>
<td></td>
<td>● tooth</td>
<td>● people</td>
</tr>
<tr>
<td></td>
<td>● vet</td>
<td>● say</td>
</tr>
<tr>
<td></td>
<td>● water</td>
<td>● like</td>
</tr>
<tr>
<td>Depressed vocabulary counts</td>
<td></td>
<td></td>
</tr>
<tr>
<td>people</td>
<td>0</td>
<td>24</td>
</tr>
<tr>
<td>know</td>
<td>0</td>
<td>18</td>
</tr>
<tr>
<td>thing</td>
<td>0</td>
<td>14</td>
</tr>
<tr>
<td>feel</td>
<td>0</td>
<td>3</td>
</tr>
<tr>
<td>time</td>
<td>0</td>
<td>3</td>
</tr>
<tr>
<td>woman</td>
<td>0</td>
<td>2</td>
</tr>
<tr>
<td>go</td>
<td>0</td>
<td>2</td>
</tr>
<tr>
<td>want</td>
<td>0</td>
<td>8</td>
</tr>
<tr>
<td>life</td>
<td>0</td>
<td>23</td>
</tr>
<tr>
<td>relationship</td>
<td>0</td>
<td>4</td>
</tr>
<tr>
<td>Control vocabulary counts</td>
<td></td>
<td></td>
</tr>
<tr>
<td>game</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>trade</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>key</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>team</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>play</td>
<td>0</td>
<td>3</td>
</tr>
<tr>
<td>player</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>shiny</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>hatch</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>thank</td>
<td>0</td>
<td>2</td>
</tr>
<tr>
<td>add</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

\(^a\)BERT: Bidirectional Encoder Representations From Transformers.
\(^b\)LM: language model.
\(^c\)TF-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 $\beta$ 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.

<table>
<thead>
<tr>
<th>Variable</th>
<th>β</th>
<th>Odds ratio</th>
<th>SE</th>
<th>P value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Depression diagnosis</td>
<td>0.163</td>
<td>1.177</td>
<td>0.035</td>
<td>&lt;.001</td>
</tr>
<tr>
<td>Time to diagnosis</td>
<td>−0.004</td>
<td>0.996</td>
<td>0.013</td>
<td>.75</td>
</tr>
<tr>
<td>Post word count</td>
<td>0.040</td>
<td>1.041</td>
<td>0.012</td>
<td>.001</td>
</tr>
<tr>
<td>Interaction (diagnosis × time)</td>
<td>0.011</td>
<td>1.011</td>
<td>0.013</td>
<td>.41</td>
</tr>
</tbody>
</table>

Table 16. Multinomial regression results for predicting whether a post is positive or negative.

<table>
<thead>
<tr>
<th>Sentiment and variable</th>
<th>β</th>
<th>Odds ratio</th>
<th>SE</th>
<th>P value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Positive</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Depression diagnosis</td>
<td>0.190</td>
<td>1.209</td>
<td>0.047</td>
<td>&lt;.001</td>
</tr>
<tr>
<td>Time to diagnosis</td>
<td>0.015</td>
<td>1.015</td>
<td>0.016</td>
<td>.37</td>
</tr>
<tr>
<td>Post word count</td>
<td>−0.070</td>
<td>0.932</td>
<td>0.019</td>
<td>&lt;.001</td>
</tr>
<tr>
<td>Interaction (diagnosis × time)</td>
<td>0.045</td>
<td>1.046</td>
<td>0.016</td>
<td>.006</td>
</tr>
<tr>
<td>Negative</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Depression diagnosis</td>
<td>0.151</td>
<td>1.163</td>
<td>0.041</td>
<td>&lt;.001</td>
</tr>
<tr>
<td>Time to diagnosis</td>
<td>−0.019</td>
<td>0.981</td>
<td>0.016</td>
<td>.24</td>
</tr>
<tr>
<td>Post word count</td>
<td>0.103</td>
<td>1.108</td>
<td>0.014</td>
<td>&lt;.001</td>
</tr>
<tr>
<td>Interaction (diagnosis × time)</td>
<td>−0.021</td>
<td>0.979</td>
<td>0.016</td>
<td>.18</td>
</tr>
</tbody>
</table>

**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_v2i1e41205_app1.xlsx ]

References

https://ai.jmir.org/2023/1/e41205

JMIR AI 2023 | vol. 2 | e41205 | p.134

(page number not for citation purposes)
24. SMHD, RSDD, and RSDD-Time Datasets. Georgetown information retrieval lab. URL: https://docs.google.com/forms/d/e/1FAIpQLScC-O3MXDd2iZSGqeRHsv1EMVBRx5NSWC0cAodsHK3tBOz_FLw/viewform [accessed 2020-11-21]


54. Simple transformers homepage. Simple Transformers. URL: https://simpletransformers.ai/ [accessed 2021-01-04]


66. The R project for statistical computing. CRAN R Project. URL: https://www.r-project.org/ [accessed 2022-07-13]


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
PMID:

©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.
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

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.

1. Vectorization
   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.

2. Calibration
   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.

3. Validation
   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).

4. Test
   A total of 100,000 patients as a test set for final analysis and presentation of the results.

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.

<table>
<thead>
<tr>
<th>Hyperparameter</th>
<th>Range</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Vector size</td>
<td>100</td>
</tr>
<tr>
<td>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.</td>
<td></td>
</tr>
<tr>
<td>2. Minimal count</td>
<td>100</td>
</tr>
<tr>
<td>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.</td>
<td></td>
</tr>
<tr>
<td>3. Window size</td>
<td>1-10</td>
</tr>
<tr>
<td>Describes how many of the neighboring codes will be considered in each training step within the 2Vec algorithm and a given sequence of codes.</td>
<td></td>
</tr>
<tr>
<td>4. Downsampling</td>
<td></td>
</tr>
<tr>
<td>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.</td>
<td></td>
</tr>
<tr>
<td>5. Epochs</td>
<td>1-20</td>
</tr>
<tr>
<td>The number of training epochs describes how many times each patient’s code sequence will be looked at to update the vectorization model.</td>
<td></td>
</tr>
<tr>
<td>6. Negative sampling</td>
<td>0-20</td>
</tr>
<tr>
<td>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.</td>
<td></td>
</tr>
<tr>
<td>7. Negative sampling exponent</td>
<td>–5 to 5</td>
</tr>
<tr>
<td>Smoothing exponent for the updates of the negative samples.</td>
<td></td>
</tr>
<tr>
<td>8. Hierarchical softmax</td>
<td>Boolean</td>
</tr>
<tr>
<td>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.</td>
<td></td>
</tr>
<tr>
<td>9. Distributed memory or distributed bag of words</td>
<td>Boolean</td>
</tr>
<tr>
<td>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.</td>
<td></td>
</tr>
<tr>
<td>10. Alpha</td>
<td>0.001-0.1</td>
</tr>
<tr>
<td>Learning rate of the neural network updates.</td>
<td></td>
</tr>
</tbody>
</table>
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 age of the patients was 45.2 years; 54.60% (5,426,481/9,937,919) of the cohort were female. The average age of the patients 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 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 6).
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.

<table>
<thead>
<tr>
<th>Characteristics</th>
<th>Values</th>
</tr>
</thead>
<tbody>
<tr>
<td>Age (years), mean (SD)</td>
<td>45.2 (24.1)</td>
</tr>
<tr>
<td>Female gender, n/N (%</td>
<td>5,426,481/9,937,919 (54.60)</td>
</tr>
<tr>
<td>Number of cases, mean (SD)</td>
<td>8.4 (6.7)</td>
</tr>
<tr>
<td>Emergency in 2019, n/N (%)</td>
<td>1,820,736/9,937,919 (18.32)</td>
</tr>
<tr>
<td>Drug cost (€), mean (SD)</td>
<td>632.1 (4383.9)</td>
</tr>
<tr>
<td>Number of codes from 2016-2018, mean (SD)</td>
<td>67.6 (92.4)</td>
</tr>
<tr>
<td>Number of codes in 2018, mean (SD)</td>
<td>34.6 (45.5)</td>
</tr>
</tbody>
</table>

*a€1=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.

### 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.

<table>
<thead>
<tr>
<th>Cluster</th>
<th>Percentage of cohort</th>
<th>Mean age (years)</th>
<th>Female, %</th>
<th>Mean number of cases</th>
<th>Emergency, %</th>
<th>Mean drug spending (€)</th>
<th>Distinctive ICD-10 codes</th>
</tr>
</thead>
<tbody>
<tr>
<td>11</td>
<td>3.8</td>
<td>4.1</td>
<td>50.4</td>
<td>4.8</td>
<td>35.2</td>
<td>69.26</td>
<td>Z00.1, Z23.8, Z27.8</td>
</tr>
<tr>
<td>8</td>
<td>1.5</td>
<td>9.4</td>
<td>35.9</td>
<td>5.7</td>
<td>27.1</td>
<td>198.01</td>
<td>F80.9, F80.0, Z00.1</td>
</tr>
<tr>
<td>6</td>
<td>1.1</td>
<td>21.7</td>
<td>49.0</td>
<td>5.3</td>
<td>21.8</td>
<td>62.77</td>
<td>H52.2, H52.0, H52.1</td>
</tr>
<tr>
<td>12</td>
<td>6.7</td>
<td>27.6</td>
<td>31.3</td>
<td>4.6</td>
<td>19.8</td>
<td>175.77</td>
<td>J06.9, A09.9, R51</td>
</tr>
<tr>
<td>1</td>
<td>1.7</td>
<td>32.0</td>
<td>99.9</td>
<td>8.4</td>
<td>28.4</td>
<td>230.47</td>
<td>Z34, N89.8, O09.3</td>
</tr>
<tr>
<td>3</td>
<td>4.0</td>
<td>33.3</td>
<td>88.5</td>
<td>7.1</td>
<td>19.1</td>
<td>323.30</td>
<td>J01.0, J45.9, J45.0</td>
</tr>
<tr>
<td>2</td>
<td>9.3</td>
<td>33.7</td>
<td>97.7</td>
<td>8.6</td>
<td>18.7</td>
<td>130.00</td>
<td>N89.8, Z30.9, Z12.9</td>
</tr>
<tr>
<td>7</td>
<td>2.6</td>
<td>44.5</td>
<td>57.1</td>
<td>9.9</td>
<td>19.0</td>
<td>431.01</td>
<td>F32.9, F32.1, F33.1</td>
</tr>
<tr>
<td>4</td>
<td>2.4</td>
<td>48.6</td>
<td>86.7</td>
<td>9.9</td>
<td>13.9</td>
<td>191.26</td>
<td>E03.9, E06.3, Z12.9</td>
</tr>
<tr>
<td>9</td>
<td>6.6</td>
<td>57.6</td>
<td>47.0</td>
<td>10.4</td>
<td>15.7</td>
<td>592.98</td>
<td>M54.1, M51.2, M54.5</td>
</tr>
<tr>
<td>10</td>
<td>3.7</td>
<td>59.3</td>
<td>37.3</td>
<td>8.4</td>
<td>11.5</td>
<td>480.11</td>
<td>I10.9, I10.90, E11.9</td>
</tr>
<tr>
<td>5</td>
<td>2.1</td>
<td>69.9</td>
<td>59.6</td>
<td>10.9</td>
<td>12.9</td>
<td>809.16</td>
<td>H52.2, H52.4, H52.0</td>
</tr>
<tr>
<td>14</td>
<td>2.6</td>
<td>74.4</td>
<td>57.3</td>
<td>8.9</td>
<td>16.0</td>
<td>1587.98</td>
<td>I10.9, I10.90, I25.1</td>
</tr>
<tr>
<td>13</td>
<td>1.3</td>
<td>80.7</td>
<td>62.9</td>
<td>8.2</td>
<td>26.6</td>
<td>1248.64</td>
<td>F03, R32, I10.9</td>
</tr>
<tr>
<td>None</td>
<td>50.7</td>
<td>50.2</td>
<td>51.0</td>
<td>9.4</td>
<td>17.9</td>
<td>908.89</td>
<td>N/A</td>
</tr>
<tr>
<td>All</td>
<td>100.0</td>
<td>45.6</td>
<td>54.5</td>
<td>8.7</td>
<td>18.7</td>
<td>654.17</td>
<td>N/A</td>
</tr>
</tbody>
</table>

*a€1=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.

<table>
<thead>
<tr>
<th>Measure</th>
<th>Linear regression</th>
<th>LightGBM Regressor</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$R^2$, %</td>
<td>Mean absolute error (€)</td>
</tr>
<tr>
<td>Age + gender</td>
<td>1.0</td>
<td>818.44</td>
</tr>
<tr>
<td>Top 100</td>
<td>2.0</td>
<td>760.55</td>
</tr>
<tr>
<td>Top 100 + age + gender</td>
<td>2.0</td>
<td>757.13</td>
</tr>
<tr>
<td>Pat2Vec</td>
<td>7.7</td>
<td>845.99</td>
</tr>
<tr>
<td>Pat2Vec + age + gender</td>
<td>7.7</td>
<td>845.98</td>
</tr>
</tbody>
</table>

*a€1=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_v21e40755_app1.pdf ]

Multimedia Appendix 2
Extended Table 2 of main manuscript. [PDF File (Adobe PDF File), 102 KB - ai_v21e40755_app2.pdf ]

References


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]


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

©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.
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-Campus1
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

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 (≤median)=low stress, 12-22=middle stress, and >22=high stress (≥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).

<table>
<thead>
<tr>
<th>Demographic characteristics</th>
<th>Values</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Continuous variables, mean (SD; range)</strong></td>
<td></td>
</tr>
<tr>
<td>Age (years)</td>
<td>42 (13.11; 18-64)</td>
</tr>
<tr>
<td>Number of persons in the household</td>
<td>3 (1.34; 1-11)</td>
</tr>
<tr>
<td>Sleep hours per night in the past 4 weeks</td>
<td>7 (1.19; 2-12)</td>
</tr>
<tr>
<td>Number of hospital nights in the past 12 months</td>
<td>1 (5.30; 0-150)</td>
</tr>
<tr>
<td>Number of sick days in the past 12 months</td>
<td>13 (38.01; 0-365)</td>
</tr>
<tr>
<td><strong>Categorical variables</strong></td>
<td></td>
</tr>
<tr>
<td>Gender (female), n (%)</td>
<td>3081 (49.6)</td>
</tr>
<tr>
<td>Marital status, n (%)</td>
<td></td>
</tr>
<tr>
<td>Married living with partner or separately from partner</td>
<td>3697 (59.5)</td>
</tr>
<tr>
<td>Single</td>
<td>1957 (31.5)</td>
</tr>
<tr>
<td>Divorced</td>
<td>376 (6.1)</td>
</tr>
<tr>
<td>Widowed</td>
<td>136 (2.2)</td>
</tr>
<tr>
<td>Provides care to someone in need or seriously ill, n (%)</td>
<td>379 (6.1)</td>
</tr>
<tr>
<td>Renting or living in own apartment/house, n (%)</td>
<td></td>
</tr>
<tr>
<td>Rented apartment or house</td>
<td>2689 (43.3)</td>
</tr>
<tr>
<td>Own apartment or house</td>
<td>3268 (52.6)</td>
</tr>
<tr>
<td>Satisfaction with living space, n (%)</td>
<td></td>
</tr>
<tr>
<td>Very satisfied or satisfied</td>
<td>5269 (84.8)</td>
</tr>
<tr>
<td>Neither satisfied nor dissatisfied</td>
<td>608 (9.8)</td>
</tr>
<tr>
<td>Dissatisfied or very dissatisfied</td>
<td>295 (4.8)</td>
</tr>
<tr>
<td>Residential area satisfaction, n (%)</td>
<td></td>
</tr>
<tr>
<td>Very satisfied or satisfied</td>
<td>5091 (81.9)</td>
</tr>
<tr>
<td>Neither satisfied nor dissatisfied</td>
<td>727 (11.7)</td>
</tr>
<tr>
<td>Dissatisfied or very dissatisfied</td>
<td>320 (5.2)</td>
</tr>
<tr>
<td>General state of health, n (%)</td>
<td></td>
</tr>
<tr>
<td>Very good or good</td>
<td>4942 (79.5)</td>
</tr>
<tr>
<td>Average</td>
<td>1134 (18.3)</td>
</tr>
<tr>
<td>Poor or very poor</td>
<td>116 (1.8)</td>
</tr>
<tr>
<td>Intake of sleeping pills in the past 4 weeks, n (%)</td>
<td></td>
</tr>
<tr>
<td>Never</td>
<td>5919 (95.3)</td>
</tr>
<tr>
<td>Less than 1 time</td>
<td>100 (1.6)</td>
</tr>
<tr>
<td>1 time or 2 times</td>
<td>73 (1.2)</td>
</tr>
<tr>
<td>3 times or more</td>
<td>86 (1.4)</td>
</tr>
<tr>
<td>Social support, n (%)</td>
<td></td>
</tr>
<tr>
<td>Low support</td>
<td>653 (10.5)</td>
</tr>
<tr>
<td>Average support</td>
<td>3082 (49.6)</td>
</tr>
<tr>
<td>Strong support</td>
<td>2451 (39.5)</td>
</tr>
<tr>
<td>Health behavior consultation in the past 12 months, n (%)</td>
<td></td>
</tr>
<tr>
<td>Has general practitioner</td>
<td>1873 (30.4)</td>
</tr>
<tr>
<td>Visited to general practitioner in the past 12 months</td>
<td>4870 (80.6)</td>
</tr>
<tr>
<td>Demographic characteristics</td>
<td>Values</td>
</tr>
<tr>
<td>-----------------------------------------------------------------</td>
<td>----------------</td>
</tr>
<tr>
<td>Visited to neurologist in the past 12 months</td>
<td>463 (7.5)</td>
</tr>
<tr>
<td><strong>Frequency of alcohol consumption, n (%)</strong></td>
<td></td>
</tr>
<tr>
<td>Never</td>
<td>744 (12.0)</td>
</tr>
<tr>
<td>1 time per month or less</td>
<td>1186 (19.1)</td>
</tr>
<tr>
<td>2-4 times per month</td>
<td>1998 (32.2)</td>
</tr>
<tr>
<td>2-3 times per week</td>
<td>1453 (23.4)</td>
</tr>
<tr>
<td>4 times per week or more</td>
<td>811 (13.1)</td>
</tr>
<tr>
<td><strong>Tobacco use, n (%)</strong></td>
<td></td>
</tr>
<tr>
<td>Yes, daily</td>
<td>1701 (27.4)</td>
</tr>
<tr>
<td>Yes, occasionally</td>
<td>433 (7)</td>
</tr>
<tr>
<td>Not anymore</td>
<td>1664 (26.8)</td>
</tr>
<tr>
<td>Never smoked</td>
<td>2400 (38.7)</td>
</tr>
<tr>
<td><strong>Comorbidities, n (%)</strong></td>
<td></td>
</tr>
<tr>
<td>Has hypertension</td>
<td>1625 (26.2)</td>
</tr>
<tr>
<td>Has diabetes</td>
<td>271 (4.4)</td>
</tr>
<tr>
<td>Has migraine</td>
<td>712 (11.5)</td>
</tr>
<tr>
<td>Has depression</td>
<td>682 (11)</td>
</tr>
<tr>
<td>Has anxiety disorder</td>
<td>327 (5.3)</td>
</tr>
<tr>
<td>Has burnout syndrome</td>
<td>292 (4.7)</td>
</tr>
<tr>
<td>Has one or more long-term chronic diseases</td>
<td>1418 (22.8)</td>
</tr>
<tr>
<td><strong>Prevention programs or sport activities, n (%)</strong></td>
<td></td>
</tr>
<tr>
<td>Participated in prevention program in the past 12 months</td>
<td>988 (15.9)</td>
</tr>
<tr>
<td>Participated in relaxation or stress management program</td>
<td>188 (3)</td>
</tr>
<tr>
<td>Participated in gymnastics, fitness, or balance sports program</td>
<td>832 (13.4)</td>
</tr>
<tr>
<td>Participated in alcohol cessation program</td>
<td>7 (0.1)</td>
</tr>
<tr>
<td>Participated in smoking cessation program</td>
<td>17 (0.3)</td>
</tr>
<tr>
<td>Participated in weight reduction or a healthy diet program</td>
<td>167 (2.7)</td>
</tr>
<tr>
<td><strong>Sports activities per week (in the past 3 months), n (%)</strong></td>
<td></td>
</tr>
<tr>
<td>No sports activity</td>
<td>1954 (31.5)</td>
</tr>
<tr>
<td>Up to 2 hours per week</td>
<td>2584 (41.6)</td>
</tr>
<tr>
<td>Regularly, 2-4 hours per week</td>
<td>990 (15.9)</td>
</tr>
<tr>
<td>Regularly, more than 4 hours per week</td>
<td>645 (10.4)</td>
</tr>
</tbody>
</table>

**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 $X = \{x_1, x_2, \ldots, x_n\}$ 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]:

$$f(x) = \sum_{k=1}^{K} \omega_k f_k(x)$$

where $f_k(x) = \omega_k$ ($q: R^m \rightarrow T, \omega R^T$) 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 $\omega$. The following regularized objective is minimized to learn the set of functions (equation 2).

$$\Omega(f) = \gamma T + \frac{1}{2} \lambda \| \omega \|^2$$

where $\Omega(f) = \gamma T + \frac{1}{2} \lambda \| \omega \|^2$, $T$ represents the model loss function, and $\Omega$ 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).

<table>
<thead>
<tr>
<th>Hyperparameter</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>learning rate</td>
<td>0.3</td>
</tr>
<tr>
<td>Estimators, n</td>
<td>1000</td>
</tr>
<tr>
<td>max_depth</td>
<td>5</td>
</tr>
<tr>
<td>Subsample</td>
<td>0.8</td>
</tr>
<tr>
<td>min_child_weight</td>
<td>3</td>
</tr>
<tr>
<td>L2 regularization term (Lambda)</td>
<td>2</td>
</tr>
<tr>
<td>colsample-bytree</td>
<td>0.7</td>
</tr>
<tr>
<td>Objective</td>
<td>multi:softmax</td>
</tr>
</tbody>
</table>

**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 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:

$$\text{they are defined as follows:}$$

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, NGBost, 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 ($X$) by calculating the impact of each feature on the prediction. The estimated Shapley values are calculated as follows (equation 9):

$$\text{estimated Shapley values are calculated as follows (equation 9):}$$

where $\hat{x}$ 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):

$$\text{The TreeSHAP interaction values estimates as follows (equation 10):}$$

where $i \neq j$, $\delta_i(S) = f_x(S \cup \{i\}) - f_x(S \cup \{j\}) - f_x(S) + f_x(S \cup \{i\} + f_x(S)$.

$M$ is the number of features, and $S$ denotes all feature subsets.

**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.

<table>
<thead>
<tr>
<th>Measure</th>
<th>XGBoost</th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Class 0</td>
<td>Class 1</td>
<td>Class 2</td>
<td>Macroaverage</td>
</tr>
<tr>
<td>AUC</td>
<td>0.83</td>
<td>0.71</td>
<td>0.89</td>
<td>0.81</td>
</tr>
<tr>
<td>Precision</td>
<td>0.73</td>
<td>0.56</td>
<td>0.58</td>
<td>0.63</td>
</tr>
<tr>
<td>Recall</td>
<td>0.80</td>
<td>0.55</td>
<td>0.37</td>
<td>0.52</td>
</tr>
<tr>
<td>Specificity</td>
<td>0.90</td>
<td>0.38</td>
<td>0.26</td>
<td>0.78</td>
</tr>
<tr>
<td>$F_1$-score</td>
<td>0.76</td>
<td>0.60</td>
<td>0.45</td>
<td>0.54</td>
</tr>
</tbody>
</table>

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.
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_v21e41868_app1.pdf ]

References


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
PMID:

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

Juliet Nabbuye Sekandi1,2, MD, MSc, DrPH; Weili Shi3, MSc; Ronghang Zhu4, MSc; Patrick Kaggwa5, BSc; Ernest Mwebaze6,7, 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, F1-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%, F1-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.

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.

<table>
<thead>
<tr>
<th>Labels</th>
<th>Description</th>
<th>Videos (N=861), n (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Positive: actual medication ingestion activities=adherence</td>
<td>• Videos show clear visibility of the face, pill, and water bottle • Patient exhibits clear action of taking pills and drinking water • Good illumination</td>
<td>405 (47)</td>
</tr>
<tr>
<td>Negative: no actual medication ingestion activities=nonadherence</td>
<td>• 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</td>
<td>92 (10)</td>
</tr>
<tr>
<td>Excluded videos</td>
<td>_a</td>
<td>364 (42.3)</td>
</tr>
<tr>
<td>Ambiguous or uncertain videos</td>
<td>• Pills not seen • Blurry faces and hands</td>
<td>157 (18.2)</td>
</tr>
<tr>
<td>Poor quality videos</td>
<td>• Poor illumination • Face of patient not seen</td>
<td>152 (17.7)</td>
</tr>
<tr>
<td>Damaged videos</td>
<td>Not reviewed</td>
<td>55 (6)</td>
</tr>
</tbody>
</table>

_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 × 1 × 1 and 3 × 3 × 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.

<table>
<thead>
<tr>
<th>Model</th>
<th>Features, n</th>
</tr>
</thead>
<tbody>
<tr>
<td>HOG*</td>
<td>16,740</td>
</tr>
<tr>
<td>Inception-v4</td>
<td>1536</td>
</tr>
<tr>
<td>3D ResNet</td>
<td>2048</td>
</tr>
<tr>
<td>3D ResNext</td>
<td>2048</td>
</tr>
<tr>
<td>Inflated 3D</td>
<td>1024</td>
</tr>
</tbody>
</table>

*HOG: histogram of oriented gradient.
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 (i.e., 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 F1-score, which are primary evaluation metrics, and sensitivity (recall), specificity, and precision (positive predictive value), which are supplementary. The F1-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 medication 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...
The performance of 3D ResNet 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).

<table>
<thead>
<tr>
<th>Feature extractor</th>
<th>Sensitivity, mean (SD)</th>
<th>Specificity, mean (SD)</th>
<th>Precision, mean (SD)</th>
<th>F₁-score, mean (SD)</th>
<th>AUC&lt;sup&gt;a&lt;/sup&gt;, mean (SD)</th>
</tr>
</thead>
<tbody>
<tr>
<td>HOG</td>
<td>90.77 (2.62)</td>
<td>27.35 (8.98)</td>
<td>85.03 (1.86)</td>
<td>87.77 (1.41)</td>
<td>0.65 (0.06)</td>
</tr>
<tr>
<td>Inception-v4</td>
<td>92.54 (3.53)</td>
<td>43.70 (8.64)</td>
<td>87.91 (1.95)</td>
<td>90.12 (1.90)</td>
<td>0.80 (0.05)</td>
</tr>
<tr>
<td>3D ResNet</td>
<td>94.57&lt;sup&gt;b&lt;/sup&gt; (2.61)</td>
<td>54.57 (6.46)</td>
<td>90.20 (1.81)</td>
<td>92.30 (1.44)</td>
<td>0.87 (0.04)</td>
</tr>
<tr>
<td>3D ResNext</td>
<td>94.17 (2.67)</td>
<td>51.74 (7.33)</td>
<td>89.62 (2.21)</td>
<td>91.81 (1.82)</td>
<td>0.85 (0.05)</td>
</tr>
<tr>
<td>Inflated 3D</td>
<td>92.94 (3.47)</td>
<td>49.78 (8.00)</td>
<td>89.08 (1.85)</td>
<td>90.94 (2.24)</td>
<td>0.82 (0.06)</td>
</tr>
</tbody>
</table>

<sup>a</sup>AUC: area under the curve.

<sup>b</sup>Italicized 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 ResNet 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 F₁-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.

<table>
<thead>
<tr>
<th>Number of dimensions</th>
<th>Sensitivity</th>
<th>Specificity</th>
<th>Precision</th>
<th>F₁-score</th>
<th>AUC&lt;sup&gt;a&lt;/sup&gt;</th>
</tr>
</thead>
<tbody>
<tr>
<td>256</td>
<td>93.09</td>
<td>51.09</td>
<td>89.39</td>
<td>91.12</td>
<td>0.83</td>
</tr>
<tr>
<td>2048</td>
<td>94.57</td>
<td>54.35</td>
<td>90.17</td>
<td>92.26</td>
<td>0.86</td>
</tr>
</tbody>
</table>

<sup>a</sup>AUC: 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.

<table>
<thead>
<tr>
<th>Method</th>
<th>Time (seconds)</th>
</tr>
</thead>
<tbody>
<tr>
<td>HOG</td>
<td>4.53</td>
</tr>
<tr>
<td>Inception-v4</td>
<td>2.38</td>
</tr>
<tr>
<td>Inflated 3D</td>
<td>0.98</td>
</tr>
<tr>
<td>3D ResNext</td>
<td>0.6</td>
</tr>
<tr>
<td>3D ResNet</td>
<td>0.54</td>
</tr>
</tbody>
</table>

*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.

<table>
<thead>
<tr>
<th>Threshold</th>
<th>Sensitivity</th>
<th>Specificity</th>
<th>Precision</th>
<th>F1-score</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.5</td>
<td>96.79</td>
<td>43.48</td>
<td>88.34</td>
<td>92.34</td>
</tr>
<tr>
<td>0.6</td>
<td>94.57</td>
<td>54.35</td>
<td>90.17</td>
<td>92.26</td>
</tr>
<tr>
<td>0.7</td>
<td>88.64</td>
<td>67.39</td>
<td>92.31</td>
<td>90.37</td>
</tr>
</tbody>
</table>

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 Nakonde, 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. JSN 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


34. Sekundi et al JMIR AI 2023 | vol. 2 | e40167 | p.176https://ai.jmir.org/2023/1/e40167


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

Meghavi Mashar1*, MBBCHIR, MA; Shreya Chawla2*, 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

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].
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 (e.g., 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 (e.g., 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 nodules 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 (e.g., 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


https://ai.jmir.org/2023/1/e42940


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]


30. Žliobaitė I, Pechenizkiy M, Gama J. An overview of concept drift applications. In: Japkowicz N, Stefanowski J, editors. First-cardiac-ultrasound-softw...


44. MD+master of engineering dual degree. Duke Biomedical Engineering. URL: https://bme.duke.edu/masters/degrees/md-meng [accessed 2022-06-24]


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

©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.
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 Wang¹,²,³,⁴, 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

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 (i.e., 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 (i.e., 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).

<table>
<thead>
<tr>
<th>Label</th>
<th>Sample, n (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>ADVANCED.LUNG.DISEASE</td>
<td>245 (55.2)</td>
</tr>
<tr>
<td>ADVANCED.HEART.DISEASE</td>
<td>117 (26.4)</td>
</tr>
<tr>
<td>CHRONIC.PAIN.FIBROMYALGIA</td>
<td>48 (10.8)</td>
</tr>
<tr>
<td>ADVANCED.CANCER</td>
<td>34 (7.7)</td>
</tr>
</tbody>
</table>

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:

$$\text{cosine similarity}(A, B) = \frac{A \cdot B}{\|A\|_2 \|B\|_2} \quad (1)$$

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 $8 \times 7 / 2 = 28$ 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 $D_{\text{train}}$ 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 \( D_{\text{train}} \), and a testing set \( D_{\text{test}} \). 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 \( x \in D_{\text{test}} \) 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, \( \text{EvalIters} \) refers to the number of averaging iterations for addressing the instability issues. In our case, \( \text{EvalIters} \) is 3.

**Textbox 1.** Algorithm 1—our proposed algorithm for Siamese neural network–style classification and evaluation for few-shot learning.

```
1: Metrics ← Zeros(\text{EvalIters}, 3);
2: for \( \text{Idx} \leftarrow 0 \) to \( \text{EvalIters} \) do
3: \( E_{\text{train}}, L_{\text{train}} \leftarrow \text{RandSubset}(D_{\text{train}}, \text{Seed} = \text{Idx}); \)
4: \( L_{\text{train}}^{2} \leftarrow \text{L2Normalize}(E_{\text{test}}); \)
5: \( L_{\text{test}}^{2} \leftarrow \text{L2Normalize}(E_{\text{train}}); \)
6: \( \text{SimilarityTable} \leftarrow \text{MatMul}(L_{\text{test}}^{2}, \text{Transpose}(L_{\text{train}}^{2})); \)
7: \( \text{LabelTable} \leftarrow \text{Zeros}(\text{Max}(L_{\text{train}}) + 1, \text{NumElements}(L_{\text{train}})); \)
8: \( \text{LabelTable}[L_{\text{train}}^{1}, \text{Arange}(\text{NumElements}(L_{\text{train}}))] \leftarrow 1; \)
9: \( \text{LabelTable} \leftarrow \text{L1Normalize}(\text{LabelTable}); \)
10: \( \text{Out} \leftarrow \text{Argmax}(\text{MatMul}(\text{SimilarityTable}, \text{Transpose}(\text{LabelTable})), \text{Dim} = 1); \)
11: \( \text{Metrics}[\text{Idx}] \leftarrow \text{ComputeMetrics}(L_{\text{test}}, \text{Out}); \)
12: end for
13: \( \text{Precision, Recall, Fscore} \leftarrow \text{Mean}(\text{Metrics}, \text{Dim} = 0). \)
```

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.

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.

\[
\text{Metric} = \left( \frac{1}{M} \sum_{i=1}^{M} \text{Metric}_i \right) / M \quad (5)
\]

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:

\[
\{\text{text}\}. \text{options are advanced cancer, advanced heart disease, advanced lung disease, chronic pain fibromyalgia. type of disease} \{\text{mask}\}
\]

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
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.**

<table>
<thead>
<tr>
<th>Approach</th>
<th>Model</th>
<th>Shots</th>
<th>Precision</th>
<th>Recall</th>
<th>F score</th>
</tr>
</thead>
<tbody>
<tr>
<td>GPT-2</td>
<td>GPT-2</td>
<td>4</td>
<td>0.63</td>
<td>0.38</td>
<td>0.42</td>
</tr>
<tr>
<td>PT-SNN</td>
<td>BERT</td>
<td>4</td>
<td>0.49</td>
<td>0.37</td>
<td>0.37</td>
</tr>
<tr>
<td>PT-SNN</td>
<td>BioBERT</td>
<td>4</td>
<td>0.53</td>
<td>0.45</td>
<td>0.46</td>
</tr>
<tr>
<td>PT-SNN</td>
<td>BioClinicalBERT</td>
<td>4</td>
<td>0.50</td>
<td>0.42</td>
<td>0.43</td>
</tr>
<tr>
<td>SOE-SNN</td>
<td>BERT</td>
<td>4</td>
<td>0.49</td>
<td>0.26</td>
<td>0.30</td>
</tr>
<tr>
<td>SOE-SNN</td>
<td>BioBERT</td>
<td>4</td>
<td>0.52</td>
<td>0.19</td>
<td>0.17</td>
</tr>
<tr>
<td>SOE-SNN</td>
<td>BioClinicalBERT</td>
<td>4</td>
<td>0.57</td>
<td>0.27</td>
<td>0.24</td>
</tr>
<tr>
<td>GPT-2</td>
<td>GPT-2</td>
<td>8</td>
<td>0.63</td>
<td>0.38</td>
<td>0.42</td>
</tr>
<tr>
<td>PT-SNN</td>
<td>BERT</td>
<td>8</td>
<td>0.62</td>
<td>0.45</td>
<td>0.47</td>
</tr>
<tr>
<td>PT-SNN</td>
<td>BioBERT</td>
<td>8</td>
<td>0.61</td>
<td>0.48</td>
<td>0.50</td>
</tr>
<tr>
<td>PT-SNN</td>
<td>BioClinicalBERT</td>
<td>8</td>
<td>0.64</td>
<td>0.44</td>
<td>0.49</td>
</tr>
<tr>
<td>SOE-SNN</td>
<td>BERT</td>
<td>8</td>
<td>0.55</td>
<td>0.43</td>
<td>0.46</td>
</tr>
<tr>
<td>SOE-SNN</td>
<td>BioBERT</td>
<td>8</td>
<td>0.61</td>
<td>0.50</td>
<td>0.53</td>
</tr>
<tr>
<td>SOE-SNN</td>
<td>BioClinicalBERT</td>
<td>8</td>
<td>0.58</td>
<td>0.32</td>
<td>0.32</td>
</tr>
<tr>
<td>GPT-2</td>
<td>GPT-2</td>
<td>16</td>
<td>0.65</td>
<td>0.38</td>
<td>0.42</td>
</tr>
<tr>
<td>PT-SNN</td>
<td>BERT</td>
<td>16</td>
<td>0.64</td>
<td>0.51</td>
<td>0.52</td>
</tr>
<tr>
<td>PT-SNN</td>
<td>BioBERT</td>
<td>16</td>
<td>0.65</td>
<td>0.55</td>
<td>0.56</td>
</tr>
<tr>
<td>PT-SNN</td>
<td>BioClinicalBERT</td>
<td>16</td>
<td>0.69</td>
<td>0.54</td>
<td>0.58</td>
</tr>
<tr>
<td>SOE-SNN</td>
<td>BERT</td>
<td>16</td>
<td>0.59</td>
<td>0.44</td>
<td>0.48</td>
</tr>
<tr>
<td>SOE-SNN</td>
<td>BioBERT</td>
<td>16</td>
<td>0.43</td>
<td>0.38</td>
<td>0.36</td>
</tr>
<tr>
<td>SOE-SNN</td>
<td>BioClinicalBERT</td>
<td>16</td>
<td>0.70</td>
<td>0.39</td>
<td>0.38</td>
</tr>
</tbody>
</table>

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-NN: 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.

References
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

©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.
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 Hsu¹², BSc, MA; Abdulaziz T Bako¹, MBBS, MPH, PhD; Thomas Potter¹, BSc, MSc, PhD; Alan P Pan¹, BSc, MSc; Gavin W Britz³⁴, MBBCHIR, MPH, MBA; Jonika Tannous¹, BA, PhD; Farhaan S Vahidy¹³⁵, 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, F₁-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

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. 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, CHARTText, 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 F1-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 (F1=0.70) achieved a similar performance as that of medical students (F1=0.73), although it was inferior to radiologists’ performance (F1=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 F1-score of 0.798, which is close to radiologists’ performances (F1=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.

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 for the model 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 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.

Table 1. Examples of text segmentation and word embedding.

<table>
<thead>
<tr>
<th>Input word</th>
<th>Word-token(s)</th>
<th>Word embedding ID(s)</th>
</tr>
</thead>
<tbody>
<tr>
<td>stroke</td>
<td>stroke</td>
<td>6625</td>
</tr>
<tr>
<td>patient</td>
<td>patient</td>
<td>5351</td>
</tr>
<tr>
<td>edema</td>
<td>(ed, #ema)</td>
<td>(5048, 14494)</td>
</tr>
<tr>
<td>hemorrhage</td>
<td>(hern, #for, #fr, #hage)</td>
<td>(23123, 1766, 1197, 19911)</td>
</tr>
</tbody>
</table>

The 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.
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 (NIHSS) at the initial encounter 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 interrviewer 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 AUROC values 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 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).

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 interrviewer agreement analysis showed an excellent agreement between the 2 annotators (0.85 % average Kappa and 97.1% agreement).
Table 2. Patient characteristics (average age and BMI are reported at imaging encounters). Italicized \( P \) values are significant.

<table>
<thead>
<tr>
<th>Characteristics</th>
<th>Head CT( ^a ) population</th>
<th>Annotation sample</th>
<th>( P ) value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Imaging notes, N</td>
<td>82,073</td>
<td>400</td>
<td></td>
</tr>
<tr>
<td>Unique patients, N</td>
<td>24,924</td>
<td>398</td>
<td></td>
</tr>
<tr>
<td>Age (years), median (Q1, Q3)</td>
<td>69.0 (58.5, 78.3)</td>
<td>68.0 (56.4, 78.1)</td>
<td>.22</td>
</tr>
<tr>
<td>Age (years), n (%)</td>
<td></td>
<td></td>
<td>.41</td>
</tr>
<tr>
<td>0-49</td>
<td>3025 (12.1)</td>
<td>57 (14.3)</td>
<td></td>
</tr>
<tr>
<td>50-59</td>
<td>3793 (15.2)</td>
<td>61 (15.3)</td>
<td></td>
</tr>
<tr>
<td>60-69</td>
<td>6149 (24.7)</td>
<td>103 (25.9)</td>
<td></td>
</tr>
<tr>
<td>( \geq 70 )</td>
<td>11,957 (48)</td>
<td>177 (44.5)</td>
<td></td>
</tr>
<tr>
<td>Gender, n (%)</td>
<td></td>
<td></td>
<td>.69</td>
</tr>
<tr>
<td>Female</td>
<td>13,439 (53.9)</td>
<td>219 (55)</td>
<td></td>
</tr>
<tr>
<td>Male</td>
<td>11,485 (46.1)</td>
<td>179 (45)</td>
<td></td>
</tr>
<tr>
<td>Race or ethnicity, n (%)</td>
<td></td>
<td></td>
<td>.22</td>
</tr>
<tr>
<td>Non-Hispanic White</td>
<td>14,028 (56.3)</td>
<td>206 (51.8)</td>
<td></td>
</tr>
<tr>
<td>Black</td>
<td>5690 (22.8)</td>
<td>102 (25.6)</td>
<td></td>
</tr>
<tr>
<td>Hispanic</td>
<td>3412 (13.7)</td>
<td>61 (15.3)</td>
<td></td>
</tr>
<tr>
<td>Asian</td>
<td>1209 (4.9)</td>
<td>16 (4)</td>
<td></td>
</tr>
<tr>
<td>Other or unknown</td>
<td>585 (2.3)</td>
<td>13 (3.3)</td>
<td></td>
</tr>
<tr>
<td>BMI (kg/m( ^2 )), median (Q1, Q3)</td>
<td>27.3 (23.7, 31.7)</td>
<td>27.3 (23.5, 31.0)</td>
<td>.59</td>
</tr>
<tr>
<td>BMI (kg/m( ^2 )), n (%)</td>
<td></td>
<td></td>
<td>.001</td>
</tr>
<tr>
<td>Underweight</td>
<td>637 (2.6)</td>
<td>13 (3.3)</td>
<td></td>
</tr>
<tr>
<td>Normal</td>
<td>6193 (24.8)</td>
<td>108 (27.1)</td>
<td></td>
</tr>
<tr>
<td>Overweight</td>
<td>6518 (26.2)</td>
<td>123 (30.9)</td>
<td></td>
</tr>
<tr>
<td>Obese</td>
<td>6610 (26.5)</td>
<td>107 (26.9)</td>
<td></td>
</tr>
<tr>
<td>Missing</td>
<td>4966 (19.9)</td>
<td>47 (11.8)</td>
<td></td>
</tr>
<tr>
<td>Insurance( ^b ), n (%)</td>
<td></td>
<td></td>
<td>.15</td>
</tr>
<tr>
<td>Medicare</td>
<td></td>
<td></td>
<td>.12</td>
</tr>
<tr>
<td>No</td>
<td>9803 (39.3)</td>
<td>142 (35.7)</td>
<td></td>
</tr>
<tr>
<td>Yes</td>
<td>15,121 (60.7)</td>
<td>256 (64.3)</td>
<td></td>
</tr>
<tr>
<td>Medicaid</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>No</td>
<td>23,793 (95.5)</td>
<td>373 (93.7)</td>
<td></td>
</tr>
<tr>
<td>Yes</td>
<td>1131 (4.5)</td>
<td>25 (6.3)</td>
<td></td>
</tr>
<tr>
<td>Commercial</td>
<td></td>
<td></td>
<td>.04</td>
</tr>
<tr>
<td>No</td>
<td>20,194 (81)</td>
<td>306 (76.9)</td>
<td></td>
</tr>
<tr>
<td>Yes</td>
<td>4730 (19)</td>
<td>92 (23.1)</td>
<td></td>
</tr>
<tr>
<td>Exchange</td>
<td></td>
<td></td>
<td>.79</td>
</tr>
<tr>
<td>No</td>
<td>24,437 (98)</td>
<td>389 (97.7)</td>
<td></td>
</tr>
<tr>
<td>Yes</td>
<td>487 (2)</td>
<td>9 (2.3)</td>
<td></td>
</tr>
<tr>
<td>Primary stroke type( ^c ), n (%)</td>
<td></td>
<td></td>
<td>&lt;.001</td>
</tr>
<tr>
<td>Subarachnoid hemorrhage</td>
<td>1307 (5.6)</td>
<td>29 (7.7)</td>
<td></td>
</tr>
<tr>
<td>Transient ischemic attack</td>
<td>7084 (30.5)</td>
<td>100 (26.5)</td>
<td></td>
</tr>
<tr>
<td>Intracerebral hemorrhage</td>
<td>2208 (9.5)</td>
<td>59 (15.6)</td>
<td></td>
</tr>
</tbody>
</table>
### Table 3. Final natural language processing model evaluation with the evaluation set (N=200) at the imaging note level.

<table>
<thead>
<tr>
<th>Stroke feature</th>
<th>Specificity</th>
<th>Precision</th>
<th>Recall</th>
<th>F1-score</th>
<th>AUROC&lt;sup&gt;a&lt;/sup&gt;, mean (95% CI)</th>
<th>Accuracy, mean (95% CI)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Midline shift</td>
<td>1</td>
<td>1</td>
<td>0.9375</td>
<td>0.9677</td>
<td>0.9730 (0.9792-1.0154)</td>
<td>0.9950 (0.9852-1.0048)</td>
</tr>
<tr>
<td>Perihematomal edema</td>
<td>0.9945</td>
<td>0.9474</td>
<td>0.9474</td>
<td>0.9474</td>
<td>0.9994 (0.9917-1.0071)</td>
<td>0.9900 (0.9762-1.0038)</td>
</tr>
<tr>
<td>White matter hyperintensity</td>
<td>0.9725</td>
<td>0.9667</td>
<td>0.956</td>
<td>0.9613</td>
<td>0.9704 (0.9452-0.9955)</td>
<td>0.9650 (0.9395-0.9905)</td>
</tr>
<tr>
<td>Lacunes</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1.0000 (1.0000-1.0000)</td>
<td>1.0000 (1.0000-1.0000)</td>
</tr>
<tr>
<td>Old stroke</td>
<td>0.9581</td>
<td>0.8056</td>
<td>0.8788</td>
<td>0.8406</td>
<td>0.9693 (0.9277-1.0110)</td>
<td>0.9450 (0.9134-0.9766)</td>
</tr>
<tr>
<td>Subacute infarct</td>
<td>0.9945</td>
<td>0.9091</td>
<td>0.5556</td>
<td>0.6897</td>
<td>0.9789 (0.9321-1.0258)</td>
<td>0.9550 (0.9263-0.9837)</td>
</tr>
<tr>
<td>Cerebral atrophy</td>
<td>0.9173</td>
<td>0.8571</td>
<td>0.9851</td>
<td>0.9167</td>
<td>0.9673 (0.9369-0.9978)</td>
<td>0.9400 (0.9071-0.9729)</td>
</tr>
<tr>
<td>Intraventricular hemorrhage</td>
<td>0.984</td>
<td>0.7273</td>
<td>0.6154</td>
<td>0.6667</td>
<td>0.9798 (0.9295-1.0338)</td>
<td>0.9600 (0.9328-0.9872)</td>
</tr>
<tr>
<td>Acute ischemia</td>
<td>0.956</td>
<td>0.6364</td>
<td>0.7778</td>
<td>0.7</td>
<td>0.9362 (0.8570-1.0154)</td>
<td>0.9400 (0.9071-0.9729)</td>
</tr>
<tr>
<td>Intracerebral hemorrhage</td>
<td>0.9665</td>
<td>0.75</td>
<td>0.8571</td>
<td>0.8</td>
<td>0.9872 (0.9532-1.0212)</td>
<td>0.9550 (0.9263-0.9837)</td>
</tr>
<tr>
<td>Subarachnoid hemorrhage</td>
<td>1</td>
<td>1</td>
<td>0.8333</td>
<td>0.9091</td>
<td>1.0000 (1.0000-1.0000)</td>
<td>0.9900 (0.9762-1.0038)</td>
</tr>
<tr>
<td>Encephalomalacia</td>
<td>1</td>
<td>1</td>
<td>0.9524</td>
<td>0.9756</td>
<td>0.9989 (0.9890-1.0088)</td>
<td>0.9950 (0.9852-1.0048)</td>
</tr>
<tr>
<td>Mass effect</td>
<td>0.9777</td>
<td>0.84</td>
<td>1</td>
<td>0.913</td>
<td>0.9952 (0.9743-1.0161)</td>
<td>0.9800 (0.9606-0.9994)</td>
</tr>
</tbody>
</table>

<sup>a</sup>AUROC: area under receiver operating characteristic curve.

### Table 4. Average natural language processing model evaluation metrics among 13 stroke features for the fine-tuned models.

<table>
<thead>
<tr>
<th>Stroke feature</th>
<th>F1-score, mean (SD)</th>
<th>AUROC&lt;sup&gt;a&lt;/sup&gt;, mean (SD)</th>
<th>Accuracy, mean (SD)</th>
</tr>
</thead>
<tbody>
<tr>
<td>HeadCT_BERT (final model)</td>
<td>0.8683 (0.1176)&lt;sup&gt;b&lt;/sup&gt;</td>
<td>0.9831 (0.0189)&lt;sup&gt;b&lt;/sup&gt;</td>
<td>0.9700 (0.0225)&lt;sup&gt;b&lt;/sup&gt;</td>
</tr>
<tr>
<td>ClinicalBERT (baseline model)</td>
<td>0.8564 (0.1173)</td>
<td>0.9786 (0.0216)</td>
<td>0.9665 (0.0237)</td>
</tr>
</tbody>
</table>

<sup>a</sup>AUROC: area under receiver operating characteristic curve.

<sup>b</sup>Italicized values denote performance of the proposed model.
### Table 5. Natural language processing (NLP) model generating structured stroke feature data sets from imaging notes.

<table>
<thead>
<tr>
<th>Characteristics</th>
<th>Head CT(^c) imaging patients(^c) (N=24924), n (%)</th>
<th>Acute ischemic stroke admission initial CT(^d) (N=6463), n (%)</th>
<th>Intracerebral hemorrhage admission initial CT(^e) (N=1243), n (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>White matter hyperintensity</td>
<td>16,014 (64.3)</td>
<td>3429 (53.1)</td>
<td>407 (32.7)</td>
</tr>
<tr>
<td>Cerebral atrophy</td>
<td>13,615 (54.6)</td>
<td>2262 (35)</td>
<td>268 (21.6)</td>
</tr>
<tr>
<td>Old stroke</td>
<td>7426 (29.8)</td>
<td>1324 (20.5)</td>
<td>91 (7.3)</td>
</tr>
<tr>
<td>Lacunes</td>
<td>6622 (26.6)</td>
<td>1386 (21.4)</td>
<td>116 (9.3)</td>
</tr>
<tr>
<td>Mass effect</td>
<td>3826 (15.4)</td>
<td>614 (9.5)</td>
<td>500 (40.2)</td>
</tr>
<tr>
<td>Intracerebral hemorrhage</td>
<td>3822 (15.3)</td>
<td>354 (5.5)</td>
<td>1096 (88.2)</td>
</tr>
<tr>
<td>Perihematomal edema</td>
<td>3600 (14.4)</td>
<td>436 (6.7)</td>
<td>623 (50.1)</td>
</tr>
<tr>
<td>Encephalomalacia</td>
<td>3453 (13.9)</td>
<td>373 (5.8)</td>
<td>50 (4)</td>
</tr>
<tr>
<td>Acute ischemia</td>
<td>3426 (13.7)</td>
<td>1173 (18.1)</td>
<td>33 (2.7)</td>
</tr>
<tr>
<td>Subacute infarct</td>
<td>2675 (10.7)</td>
<td>841 (13)</td>
<td>28 (2.3)</td>
</tr>
<tr>
<td>Subarachnoid hemorrhage</td>
<td>2179 (8.7)</td>
<td>132 (2)</td>
<td>245 (19.7)</td>
</tr>
<tr>
<td>Midline shift</td>
<td>1908 (7.7)</td>
<td>184 (2.8)</td>
<td>345 (27.8)</td>
</tr>
<tr>
<td>Intraventricular hemorrhage</td>
<td>1409 (5.7)</td>
<td>37 (0.6)</td>
<td>405 (32.6)</td>
</tr>
<tr>
<td>Severe stroke features(^f)</td>
<td>5146 (20.6)</td>
<td>901 (13.9)</td>
<td>845 (68)</td>
</tr>
</tbody>
</table>

\(^a\)Our 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.

\(^b\)CT: computed tomography.

\(^c\)The 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.

\(^f\)Severe 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 features. We 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.

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


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

©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

https://ai.jmir.org/2023/1/e42884 JMIR AI 2023 | vol. 2 | e42884 | p.207

(page number not for citation purposes)
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.
Natural Language Processing for Clinical Laboratory Data Repository Systems: Implementation and Evaluation for Respiratory Viruses

Elham Dolatabadi1,2,3, BSc, MSc, PhD; Branson Chen4, BHS, MSc; Sarah A Buchan3,4,5,6, MSc, PhD; Alex Marchand Austin4, BSc, MSc; Mahmoud Azimaei3,4, BSc; Allison McGeer3,6,7,8, MSc, MD, PhD; Samira Mubareka8,9, MD; Jeffrey C Kwong4,5,6,10,11, 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 F1-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*

**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, adenoviruses, seasonal coronaviruses, enterovirus/rhinoviruses, parainfluenza viruses, HMV, and bocavirus.

<table>
<thead>
<tr>
<th>Viruses</th>
<th>Mention counts(^d), n (%)</th>
<th>Tested(^e), n (%)</th>
<th>Detected(^f), n (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Adenovirus</td>
<td>21,614 (7)</td>
<td>45 (6)</td>
<td>2 (1)</td>
</tr>
<tr>
<td>Bocavirus</td>
<td>5112 (2)</td>
<td>96 (13)</td>
<td>5 (3)</td>
</tr>
<tr>
<td>Coronavirus (seasonal)</td>
<td>9128 (3)</td>
<td>95 (13)</td>
<td>9 (5)</td>
</tr>
<tr>
<td>Any influenza</td>
<td>49,282 (16)</td>
<td>78 (11)</td>
<td>35 (20)</td>
</tr>
<tr>
<td>Influenza A</td>
<td>44,753 (15)</td>
<td>80 (11)</td>
<td>30 (18)</td>
</tr>
<tr>
<td>Influenza A H1</td>
<td>6797 (2)</td>
<td>N/A(^g)</td>
<td>17 (10)</td>
</tr>
<tr>
<td>Influenza A H3</td>
<td>9929 (3)</td>
<td>N/A</td>
<td>18 (10)</td>
</tr>
<tr>
<td>Influenza B</td>
<td>40,840 (13)</td>
<td>78 (11)</td>
<td>12 (7)</td>
</tr>
<tr>
<td>Enterovirus/rhinovirus</td>
<td>13,262 (4)</td>
<td>92 (13)</td>
<td>19 (11)</td>
</tr>
<tr>
<td>HMV</td>
<td>21,194 (7)</td>
<td>46 (6)</td>
<td>3 (2)</td>
</tr>
<tr>
<td>Parainfluenza</td>
<td>21,584 (7)</td>
<td>46 (6)</td>
<td>4 (2)</td>
</tr>
<tr>
<td>Any RSV</td>
<td>38,080 (12)</td>
<td>68 (9)</td>
<td>11 (6)</td>
</tr>
<tr>
<td>RSV A</td>
<td>11,227 (4)</td>
<td>N/A</td>
<td>2 (1)</td>
</tr>
<tr>
<td>RSV B</td>
<td>11,094 (4)</td>
<td>N/A</td>
<td>3 (2)</td>
</tr>
<tr>
<td>Total</td>
<td>303,896 (100)</td>
<td>724 (100)</td>
<td>170 (100)</td>
</tr>
</tbody>
</table>

\(^a\)RSV: respiratory syncytial virus.
\(^b\)HMV: human metapneumovirus.
\(^c\)The 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.
\(^d\)Represents 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.
\(^e\)Represents the proportion of mentions flagged as tested by the parser.
\(^f\)Represents 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 e a result) and then flagged it as detected if the result is positive. Detected is a subset of the tested.
\(^g\)N/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

The tabular expected output:

- Bovacivirus is tested.
- Influenza A is tested.
- Influenza B is tested.
- Enterovirus/Rhinovirus is tested.
- CMV is tested.
- RSV is tested.
- Parainfluenza is tested.
- Influenza A is detected.
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 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.

<table>
<thead>
<tr>
<th>Cohorts</th>
<th>Sequences(^a), n (%)</th>
<th>Any influenza virus(^b)</th>
<th>Any RSV(^c) virus</th>
<th>Any virus</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Detected, n (%)</td>
<td>Tested, n (%)</td>
<td>Detected, n (%)</td>
<td>Tested, n (%)</td>
</tr>
<tr>
<td>Development set (2009-2018)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Training set</td>
<td>60,471 (69)</td>
<td>13,792 (16)</td>
<td>35,292 (40)</td>
<td>3959 (4)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>27,196 (31)</td>
</tr>
<tr>
<td>Internal test set</td>
<td>6719 (8)</td>
<td>1604 (2)</td>
<td>3941 (4)</td>
<td>428 (0.5)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>3009 (3)</td>
</tr>
<tr>
<td>Out-of-time test sets</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Pre–COVID-19 (2019)</td>
<td>15,908 (18)</td>
<td>3019 (3)</td>
<td>6903 (8)</td>
<td>706 (0.8)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>5957 (7)</td>
</tr>
<tr>
<td>Post–COVID-19 (2020)</td>
<td>100 (0.01)</td>
<td>N/A(^d)</td>
<td>11 (0.01)</td>
<td>&lt;6 (0.006)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>External test set (2009-2018)</td>
<td>4213 (5)</td>
<td>864 (1)</td>
<td>3020 (34)</td>
<td>261 (0.2)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>2546 (3)</td>
</tr>
</tbody>
</table>

\(^a\)Represents 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.

\(^b\)Detected and tested represent the aggregation of the proportion of any mentions of the virus terms from the total unique sequences in the data set.

\(^c\)RSV: respiratory syncytial virus.

\(^d\)N/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.

<table>
<thead>
<tr>
<th>Variables</th>
<th>Internal test set</th>
<th>Out-of-time test set</th>
<th>External test set</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>(Post–COVID-19)</td>
<td>(Pre–COVID-19)</td>
<td></td>
</tr>
<tr>
<td>Fine-grained microaveraged, mean (SD)</td>
<td>97.3 (0.25)</td>
<td>94.31 (0.59)</td>
<td>60.45 (7.99)</td>
</tr>
<tr>
<td><strong>Coarse-grained any influenza virus, mean (SD)</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Detected</td>
<td>97.64 (0.28)</td>
<td>94.47 (1.04)</td>
<td>N/A</td>
</tr>
<tr>
<td>Tested</td>
<td>98.71 (0.15)</td>
<td>97.26 (0.45)</td>
<td>69.8 (4.43)</td>
</tr>
<tr>
<td><strong>Coarse-grained any RSV, mean (SD)</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Detected</td>
<td>90.94 (1.7)</td>
<td>81.56 (3.63)</td>
<td>48.33 (44.76)</td>
</tr>
<tr>
<td>Tested</td>
<td>98.16 (0.34)</td>
<td>96.18 (0.95)</td>
<td>95.6 (5.69)</td>
</tr>
<tr>
<td><strong>Coarse-grained any virus, mean (SD)</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Detected</td>
<td>95.01 (1)</td>
<td>92.31 (1.59)</td>
<td>31.71 (9.44)</td>
</tr>
<tr>
<td>Tested</td>
<td>98.4 (0.17)</td>
<td>96.3 (0.35)</td>
<td>75.87 (4.82)</td>
</tr>
</tbody>
</table>

a The 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.

b Detected and tested represent the aggregation of the proportion of any mentions of the virus terms from the total unique sequences in the data set.

c N/A: not applicable.

d 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 difficult.
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


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

©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.
Automated Identification of Aspirin-Exacerbated Respiratory Disease Using Natural Language Processing and Machine Learning: Algorithm Development and Evaluation Study

Thanai Pongdee, MD; Nicholas B Larson, MS, PhD; Rohit Divekar, MBBS, PhD; Suzette J Bielinski, PhD; Hongfang Liu, PhD; Sungrim Moon, 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.

https://ai.jmir.org/2023/1/e44191

doi: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 polyph 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:
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).

<table>
<thead>
<tr>
<th>AERD Feature</th>
<th>Train, n (%)</th>
<th>Test, n (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>AERD</td>
<td>103 (52)</td>
<td>60 (30)</td>
</tr>
<tr>
<td>Asthma</td>
<td>192 (96)</td>
<td>82 (41)</td>
</tr>
<tr>
<td>NSAID^a allergy</td>
<td>98 (49)</td>
<td>121 (61)</td>
</tr>
<tr>
<td>Nasal polyps</td>
<td>175 (88)</td>
<td>192 (96)</td>
</tr>
<tr>
<td>Chronic sinusitis</td>
<td>182 (91)</td>
<td>180 (90)</td>
</tr>
<tr>
<td>LAB^b</td>
<td>93 (47)</td>
<td>179 (90)</td>
</tr>
<tr>
<td>Documented no-NSAID allergy</td>
<td>70 (35)</td>
<td>101 (51)</td>
</tr>
</tbody>
</table>

^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.

<table>
<thead>
<tr>
<th>Rule</th>
<th>Sequential rules</th>
<th>AERD</th>
<th>Case, n</th>
<th>Correct, n</th>
<th>Error, n</th>
<th>Confidence (%)&lt;sup&gt;a&lt;/sup&gt;</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>AERD≤3.5, NSAID allergy≤2.5, Chronic Sinusitis≤6.5, and then documented non-NSAID allergy≤0.5</td>
<td>No</td>
<td>30</td>
<td>12</td>
<td>0</td>
<td>40</td>
</tr>
<tr>
<td>2</td>
<td>AERD≤3.5, NSAID allergy≤2.5, Chronic Sinusitis≤6.5, and then documented non-NSAID allergy&gt;0.5</td>
<td>No</td>
<td>9</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>3</td>
<td>AERD≤3.5, NSAID allergy≤2.5, Chronic Sinusitis&gt;6.5, and then documented non-NSAID allergy≤0.5</td>
<td>No</td>
<td>43</td>
<td>40</td>
<td>0</td>
<td>93</td>
</tr>
<tr>
<td>4</td>
<td>AERD≤3.5, NSAID allergy≤2.5, Chronic Sinusitis&gt;6.5, and then documented non-NSAID allergy&gt;0.5</td>
<td>No</td>
<td>4</td>
<td>3</td>
<td>0</td>
<td>75</td>
</tr>
<tr>
<td>5</td>
<td>AERD≤3.5, NSAID allergy&gt;2.5, and then Chronic Sinusitis≤9.0</td>
<td>Yes</td>
<td>10</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>6</td>
<td>AERD≤3.5, NSAID allergy&gt;2.5, and then Chronic Sinusitis&gt;9.0</td>
<td>Yes</td>
<td>74</td>
<td>1</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>7</td>
<td>AERD&gt;3.5, NSAID allergy≤1.5, and then LAB&lt;0.5</td>
<td>Yes</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>8</td>
<td>AERD&gt;3.5, NSAID allergy≤1.5, and then LAB&gt;0.5</td>
<td>No</td>
<td>2</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>9</td>
<td>AERD&gt;3.5, NSAID allergy&gt;1.5</td>
<td>Yes</td>
<td>2</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>10</td>
<td>Others</td>
<td>Yes</td>
<td>16</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td></td>
<td>No</td>
<td>10</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

N/A<sup>e</sup> The cases were not identified according to the original intended sequential rule; instead, a different sequence rule was used.

<sup>a</sup>Confidence = 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.

<sup>b</sup>NSAID allergy refers to diagnosis of allergy to a nonsteroidal anti-inflammatory drug (NSAID).

<sup>c</sup>Chronic sinusitis refers to diagnosis of chronic sinusitis.

<sup>d</sup>LAB refers to a prior abnormally elevated urine leukotriene E4 level.

<sup>e</sup>N/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.

<table>
<thead>
<tr>
<th>Data set</th>
<th>Sensitivity (%; 95% CI)</th>
<th>Specificity (%; 95% CI)</th>
<th>Positive predictive value (%; 95% CI)</th>
<th>Negative predictive value (%; 95% CI)</th>
<th>Accuracy (%; 95% CI)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Train</td>
<td>88.00 (79.98-93.64)</td>
<td>97.00 (91.48-99.38)</td>
<td>96.70 (90.67-99.31)</td>
<td>88.99 (81.56-94.18)</td>
<td>92.50 (87.93-95.74)</td>
</tr>
<tr>
<td>Test</td>
<td>80.00 (70.82-87.33)</td>
<td>88.00 (79.98-93.640</td>
<td>86.96 (78.32-93.07)</td>
<td>81.48 (72.86-88.31)</td>
<td>84.00 (78.17-88.79)</td>
</tr>
</tbody>
</table>

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


**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
Extractive Clinical Question-Answering With Multianswer and Multifocus Questions: Data Set Development and Evaluation Study

Sungrim Moon, PhD; Huan He, PhD; Heling Jia, MD; Hongfang Liu, PhD; Jungwei Wilfred Fan, 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 F₁-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

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 (e.g., 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.
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.

<table>
<thead>
<tr>
<th>Category in the n2c2 corpus</th>
<th>Example</th>
<th>Conversion strategy</th>
<th>Entries, n</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 Drug, no Reason</td>
<td>Mirtazapine 15 mg PO QHS(^b) (only the drug is mentioned but no reason is documented)</td>
<td>Make an unanswerable QA(^c) entry</td>
<td>46,278</td>
</tr>
<tr>
<td>1 Drug, 1 Reason</td>
<td>The patient received morphine for pain as needed</td>
<td>Make a 1-to-1 QA entry</td>
<td>28,224(^d)</td>
</tr>
<tr>
<td>N Drugs, 1 Reason</td>
<td>Hypertension: severely elevated blood pressure. Started amiodipine, metoprolol, and isosorbide.</td>
<td>Break into N separate 1-to-1 relations and make each a 1-to-1 QA entry</td>
<td>N/A(^e)</td>
</tr>
<tr>
<td>1 Drug, N Reasons</td>
<td>Albuterol sulfate 90 mcg… Puff Inhalation Q4H(^f) for sob or wheeze.</td>
<td>List the N reasons under the answer block to form a 1-to-N QA entry</td>
<td>22,437(^g)</td>
</tr>
<tr>
<td>M Drug, N Reasons</td>
<td>Left frontoparietal stroke - maintained on ASA(^h) and plavix …. Hx of CVA(^i): restarted ASA/Plavix per the GI(^j) team’s recommendation.</td>
<td>List the N reasons under answer block to form an M-to-N QA entry</td>
<td>N/A</td>
</tr>
</tbody>
</table>

\(^a\)n2c2: National NLP (natural language processing) Clinical Challenges.

\(^b\)PO QHS: one pill to be taken orally at bedtime.

\(^c\)QA: question-answering.

\(^d\)28,224 entries in total for the 1-drug-1-reason and N-drugs-1-reason categories together in the corpus.

\(^e\)N/A: not applicable.

\(^f\)Q4H: every 4 hours.

\(^g\)22,437 entries in total for the 1-drug-N-reasons and M-drug-N-reasons categories in together in the corpus.

\(^h\)ASA: acetylsalicylic acid (aspirin).

\(^i\)Hx of CVA: history of cerebrovascular accident.

\(^j\)GI: gastrointestinal.

Table 2. Unique answers among answerable questions.

<table>
<thead>
<tr>
<th>Frequency</th>
<th>Unique answers, n (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>28,224 (75)</td>
</tr>
<tr>
<td>2</td>
<td>6804 (18)</td>
</tr>
<tr>
<td>3</td>
<td>1530 (4)</td>
</tr>
<tr>
<td>≥4</td>
<td>954 (3)</td>
</tr>
</tbody>
</table>

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.

```json
{
  "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).

<table>
<thead>
<tr>
<th>Distance (be sentence) between the question and answer items</th>
<th>QA entries, n</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>32,409</td>
</tr>
<tr>
<td>1</td>
<td>8154</td>
</tr>
<tr>
<td>2</td>
<td>2646</td>
</tr>
<tr>
<td>3</td>
<td>1188</td>
</tr>
<tr>
<td>4</td>
<td>405</td>
</tr>
<tr>
<td>5</td>
<td>153</td>
</tr>
<tr>
<td>6</td>
<td>81</td>
</tr>
<tr>
<td>7</td>
<td>72</td>
</tr>
<tr>
<td>8</td>
<td>27</td>
</tr>
<tr>
<td>9</td>
<td>0</td>
</tr>
<tr>
<td>10</td>
<td>0</td>
</tr>
<tr>
<td>11</td>
<td>0</td>
</tr>
<tr>
<td>12</td>
<td>9</td>
</tr>
<tr>
<td>13</td>
<td>9</td>
</tr>
<tr>
<td>14</td>
<td>0</td>
</tr>
<tr>
<td>15</td>
<td>0</td>
</tr>
<tr>
<td>16</td>
<td>9</td>
</tr>
</tbody>
</table>

*QA: 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.

<table>
<thead>
<tr>
<th>Pattern</th>
<th>Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>Reason … (being)? [received</td>
<td>started</td>
</tr>
<tr>
<td>Drug … [pmPRN] (as needed for)? Reason</td>
<td>18</td>
</tr>
<tr>
<td>Drug … (was)? [attempted</td>
<td>given</td>
</tr>
<tr>
<td>Reason … (was)? [managed</td>
<td>treated</td>
</tr>
</tbody>
</table>

F1-Score of the Baseline EQA Solution

The performance in determining the F1-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 masking 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

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]

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


19. n2c2 Data Upload: Community generated annotations. DBMI Data Portal. URL: https://portal.dbmi.hms.harvard.edu/projects/n2c2-du/ [accessed 2023-05-29]


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 Khriguian², MD; Marc David², MD; John Kildlea¹, 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 point–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://ai.jmir.org/2023/1/e44779
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 by which 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.

<table>
<thead>
<tr>
<th>Characteristics</th>
<th>Training and validation set (n=121)</th>
<th>Test set (n=55)</th>
<th>P value(^a)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gender, n (%)</td>
<td></td>
<td></td>
<td>N/A(^b)</td>
</tr>
<tr>
<td>Female</td>
<td>56 (46)</td>
<td>25 (45)</td>
<td></td>
</tr>
<tr>
<td>Male</td>
<td>65 (54)</td>
<td>30 (55)</td>
<td></td>
</tr>
<tr>
<td>Age (years), mean (SD)</td>
<td>.99</td>
<td>.72</td>
<td></td>
</tr>
<tr>
<td>Female</td>
<td>63 (14)</td>
<td>64 (12)</td>
<td></td>
</tr>
<tr>
<td>Male</td>
<td>67 (14)</td>
<td>64 (13)</td>
<td></td>
</tr>
<tr>
<td>Primary cancer type, n (%)</td>
<td>.06</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Lung</td>
<td>32 (26)</td>
<td>20 (36)</td>
<td></td>
</tr>
<tr>
<td>Breast</td>
<td>23 (19)</td>
<td>11 (20)</td>
<td></td>
</tr>
<tr>
<td>Prostate</td>
<td>19 (16)</td>
<td>5 (9)</td>
<td></td>
</tr>
<tr>
<td>Multiple myeloma</td>
<td>8 (7)</td>
<td>6 (11)</td>
<td></td>
</tr>
<tr>
<td>Renal cell carcinoma</td>
<td>7 (6)</td>
<td>2 (4)</td>
<td></td>
</tr>
<tr>
<td>Other and unknown</td>
<td>64 (53)</td>
<td>31 (56)</td>
<td></td>
</tr>
<tr>
<td>Bone metastasis lesions, n (%)</td>
<td>.42</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Lytic</td>
<td>220 (52)</td>
<td>76 (47)</td>
<td></td>
</tr>
<tr>
<td>Blastic</td>
<td>122 (29)</td>
<td>57 (35)</td>
<td></td>
</tr>
<tr>
<td>Mix</td>
<td>81 (19)</td>
<td>30 (18)</td>
<td></td>
</tr>
<tr>
<td>Pain label, n (%)</td>
<td>N/A</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Pain</td>
<td>357 (84)</td>
<td>136 (83)</td>
<td></td>
</tr>
<tr>
<td>No pain</td>
<td>66 (16)</td>
<td>27 (17)</td>
<td></td>
</tr>
</tbody>
</table>

\(^a\)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.

\(^b\)N/A: not applicable.

\(^c\)The 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 API=0), and VDP=“pain” (if API>0). 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.”
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 an 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.

<table>
<thead>
<tr>
<th>Nested spherical (SP) ROIs with Hounsfield units (HUs) intensity thresholds (HUs&gt;0):</th>
</tr>
</thead>
<tbody>
<tr>
<td>• SP7 (diameter 7 mm)</td>
</tr>
<tr>
<td>• SP10 (diameter 10 mm)</td>
</tr>
<tr>
<td>• SP15 (diameter 15 mm)</td>
</tr>
<tr>
<td>• SP20 (diameter 20 mm)</td>
</tr>
<tr>
<td>• SP30 (diameter 30 mm)</td>
</tr>
<tr>
<td>• SP50 (diameter 50 mm)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Ensemble (EN) ROIs:</th>
</tr>
</thead>
<tbody>
<tr>
<td>• EN3 (ROI SP7+SP10+SP15)</td>
</tr>
<tr>
<td>• EN6 (ROI SP7+SP10+SP15+SP20+SP30+SP50)</td>
</tr>
</tbody>
</table>

**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 F1-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; ROE: 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. 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.

<table>
<thead>
<tr>
<th>Region of interest</th>
<th>Training set</th>
<th>Test set</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>RF</td>
<td>GPR</td>
</tr>
<tr>
<td>EN3</td>
<td>98.3</td>
<td>98.1</td>
</tr>
<tr>
<td>EN6</td>
<td>98.1</td>
<td>98.3</td>
</tr>
</tbody>
</table>

$F_1$-scores

<table>
<thead>
<tr>
<th>Region of interest</th>
<th>Training set</th>
<th>Test set</th>
</tr>
</thead>
<tbody>
<tr>
<td>EN3</td>
<td>90.0</td>
<td>90.5</td>
</tr>
<tr>
<td>EN6</td>
<td>93.0</td>
<td>93.0</td>
</tr>
</tbody>
</table>

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].

<table>
<thead>
<tr>
<th></th>
<th>Accuracy</th>
<th>Precision</th>
<th>Sensitivity</th>
<th>Specificity</th>
<th>$F_1$-score</th>
<th>AUC$^a$</th>
</tr>
</thead>
<tbody>
<tr>
<td>This study (training set)</td>
<td>92.4</td>
<td>93.2</td>
<td>92.4</td>
<td>86.4</td>
<td>91.6</td>
<td>94.0</td>
</tr>
<tr>
<td>This study (test set)</td>
<td>81.0</td>
<td>67.9</td>
<td>59.2</td>
<td>85.3</td>
<td>69.5</td>
<td>82.5</td>
</tr>
<tr>
<td>This study (training set); using manual pain scores</td>
<td>94.2</td>
<td>94.8</td>
<td>98.7</td>
<td>89.7</td>
<td>94.4</td>
<td>98.1</td>
</tr>
<tr>
<td>This study (test set); using manual pain scores</td>
<td>83.5</td>
<td>64.9</td>
<td>64.7</td>
<td>85.7</td>
<td>68.0</td>
<td>82.3</td>
</tr>
<tr>
<td>Wakabayashi et al [13] (training test only)</td>
<td>73.9</td>
<td>—$^b$</td>
<td>71.0</td>
<td>86.0</td>
<td>—</td>
<td>82.0</td>
</tr>
</tbody>
</table>

$^a$AUC: area under the receiver operating characteristic curve.

$^b$Not 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.

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.

References


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
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 Lee¹*, PhD; Zongzhi Liu¹*, 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 Schadt¹, 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% F₁-scores on different GGO features. We deployed this GGO NLP model to extract and structure GGO features and inform the longitudinal trajectory of GGO status from large-scale radiology notes. 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

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, and 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 solidity, 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 solidity 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 (LLL), 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 $2 \times PPV \times sensitivity/(PPV + sensitivity)$. 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 Sanai/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 cohort.

<table>
<thead>
<tr>
<th>Variables</th>
<th>Overall (N=4521), n (%)</th>
<th>GGO cohort persistency</th>
<th>Nonpersistent GGO (n=1252), n (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Persistent GGO (n=3269), n (%)</td>
<td></td>
</tr>
<tr>
<td>Gender</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Female</td>
<td>2431 (53.77)</td>
<td>1790 (54.76)</td>
<td>641 (51.20)</td>
</tr>
<tr>
<td>Male</td>
<td>2090 (46.23)</td>
<td>1479 (45.24)</td>
<td>611 (48.80)</td>
</tr>
<tr>
<td>Race</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>White</td>
<td>2394 (52.95)</td>
<td>1700 (52)</td>
<td>694 (55.43)</td>
</tr>
<tr>
<td>Other</td>
<td>791 (17.50)</td>
<td>603 (18.45)</td>
<td>188 (15.02)</td>
</tr>
<tr>
<td>Black or African American</td>
<td>722 (15.97)</td>
<td>530 (16.21)</td>
<td>192 (15.34)</td>
</tr>
<tr>
<td>Unknown</td>
<td>363 (8.03)</td>
<td>244 (7.46)</td>
<td>119 (9.50)</td>
</tr>
<tr>
<td>Asian</td>
<td>165 (3.65)</td>
<td>139 (4.25)</td>
<td>26 (2.08)</td>
</tr>
<tr>
<td>Native Hawaiian or other Pacific Islander</td>
<td>83 (1.84)</td>
<td>50 (1.53)</td>
<td>33 (2.64)</td>
</tr>
<tr>
<td>American Indian or Alaska Native</td>
<td>3 (0.07)</td>
<td>3 (0.09)</td>
<td>0 (0)</td>
</tr>
<tr>
<td>Ethnicity</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Not Hispanic or Latino</td>
<td>2442 (54.01)</td>
<td>1864 (57.02)</td>
<td>578 (46.17)</td>
</tr>
<tr>
<td>Unknown</td>
<td>1492 (33)</td>
<td>955 (29.21)</td>
<td>537 (42.89)</td>
</tr>
<tr>
<td>Hispanic or Latino</td>
<td>519 (11.48)</td>
<td>399 (12.21)</td>
<td>120 (9.58)</td>
</tr>
<tr>
<td>Not reported</td>
<td>68 (1.50)</td>
<td>51 (1.56)</td>
<td>17 (1.36)</td>
</tr>
<tr>
<td>Smoking status</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>No record of smoking</td>
<td>2304 (50.96)</td>
<td>1557 (47.63)</td>
<td>747 (59.66)</td>
</tr>
<tr>
<td>Former smoker</td>
<td>1287 (28.47)</td>
<td>996 (30.47)</td>
<td>291 (23.24)</td>
</tr>
<tr>
<td>Never smoker</td>
<td>511 (11.30)</td>
<td>395 (12.08)</td>
<td>116 (9.27)</td>
</tr>
<tr>
<td>Smoker</td>
<td>414 (9.16)</td>
<td>317 (9.70)</td>
<td>97 (7.75)</td>
</tr>
<tr>
<td>Passive smoker</td>
<td>5 (0.11)</td>
<td>4 (0.12)</td>
<td>1 (0.08)</td>
</tr>
<tr>
<td>Comorbidities</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>History of COPD&lt;sup&gt;c&lt;/sup&gt;</td>
<td>604 (13.36)</td>
<td>444 (13.58)</td>
<td>160 (12.78)</td>
</tr>
<tr>
<td>History of heart disease</td>
<td>1297 (28.69)</td>
<td>924 (28.27)</td>
<td>373 (29.79)</td>
</tr>
<tr>
<td>History of chronic kidney disease</td>
<td>340 (7.52)</td>
<td>262 (8.01)</td>
<td>78 (6.23)</td>
</tr>
<tr>
<td>History of NMSC&lt;sup&gt;d&lt;/sup&gt;</td>
<td>36 (0.80)</td>
<td>27 (0.83)</td>
<td>9 (0.72)</td>
</tr>
<tr>
<td>History of any cancer except NMSC</td>
<td>3086 (68.26)</td>
<td>2189 (66.96)</td>
<td>897 (71.65)</td>
</tr>
<tr>
<td>Family history</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Family history of lung cancer</td>
<td>8 (0.18)</td>
<td>7 (0.21)</td>
<td>1 (0.08)</td>
</tr>
<tr>
<td>Family history of any cancer</td>
<td>79 (1.75)</td>
<td>63 (1.93)</td>
<td>16 (1.28)</td>
</tr>
</tbody>
</table>

<sup>a</sup>GGO: ground-glass opacity.
<sup>b</sup>Each patient can have more than 1 comorbidity.
<sup>c</sup>COPD: chronic obstructive pulmonary disease.
<sup>d</sup>NMSC: 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\textsuperscript{a} pipeline.

<table>
<thead>
<tr>
<th>Semantic</th>
<th>Right\textsuperscript{b}</th>
<th>Predict\textsuperscript{c}</th>
<th>Gold\textsuperscript{d}</th>
<th>Precision</th>
<th>Recall</th>
<th>$F_1$-score</th>
</tr>
</thead>
<tbody>
<tr>
<td>GGO\textsuperscript{e} term: pure GGO</td>
<td>986</td>
<td>987</td>
<td>989</td>
<td>0.99</td>
<td>1</td>
<td>0.99</td>
</tr>
<tr>
<td>GGO term: part-solid GGO</td>
<td>145</td>
<td>146</td>
<td>146</td>
<td>0.99</td>
<td>0.99</td>
<td>0.99</td>
</tr>
<tr>
<td>GGO solidity</td>
<td>99</td>
<td>99</td>
<td>100</td>
<td>1</td>
<td>0.99</td>
<td>0.99</td>
</tr>
<tr>
<td>GGO shape/margin</td>
<td>144</td>
<td>151</td>
<td>144</td>
<td>0.95</td>
<td>1</td>
<td>0.98</td>
</tr>
<tr>
<td>GGO size</td>
<td>653</td>
<td>659</td>
<td>667</td>
<td>0.99</td>
<td>0.98</td>
<td>0.98</td>
</tr>
<tr>
<td>GGO quantity</td>
<td>154</td>
<td>156</td>
<td>160</td>
<td>0.99</td>
<td>0.96</td>
<td>0.97</td>
</tr>
<tr>
<td>GGO status change: better</td>
<td>46</td>
<td>46</td>
<td>46</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>GGO status change: worsen</td>
<td>107</td>
<td>107</td>
<td>110</td>
<td>1</td>
<td>0.97</td>
<td>0.99</td>
</tr>
<tr>
<td>GGO status change: stable</td>
<td>510</td>
<td>535</td>
<td>572</td>
<td>0.95</td>
<td>0.89</td>
<td>0.92</td>
</tr>
<tr>
<td>Potential GGO cause: infectious/inflammatory</td>
<td>146</td>
<td>147</td>
<td>148</td>
<td>0.99</td>
<td>0.99</td>
<td>0.99</td>
</tr>
<tr>
<td>Potential GGO cause: neoplasm</td>
<td>121</td>
<td>122</td>
<td>132</td>
<td>0.99</td>
<td>0.92</td>
<td>0.95</td>
</tr>
<tr>
<td>Potential GGO cause: others</td>
<td>71</td>
<td>73</td>
<td>76</td>
<td>0.97</td>
<td>0.95</td>
<td>0.95</td>
</tr>
<tr>
<td>GGO location</td>
<td>1164</td>
<td>1220</td>
<td>1270</td>
<td>0.95</td>
<td>0.92</td>
<td>0.93</td>
</tr>
<tr>
<td>Temporal</td>
<td>1650</td>
<td>1700</td>
<td>1650</td>
<td>0.97</td>
<td>1</td>
<td>0.99</td>
</tr>
</tbody>
</table>

\textsuperscript{a}NLP: natural language processing.
\textsuperscript{b}The number of accurately extracted entities based on the gold standard.
\textsuperscript{c}The number of entities predicted from the NLP pipeline.
\textsuperscript{d}Manually annotated entity by annotators.
\textsuperscript{e}GGO: 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>
<thead>
<tr>
<th>GGO attributes</th>
<th>Patients (N=4521), n (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pure GGO</td>
<td>4432 (98)</td>
</tr>
<tr>
<td>Part solid GGO</td>
<td>505 (11)</td>
</tr>
<tr>
<td><strong>Location</strong></td>
<td></td>
</tr>
<tr>
<td>Bilateral/both</td>
<td>2277 (63.5)</td>
</tr>
<tr>
<td>Left</td>
<td>438 (12.2)</td>
</tr>
<tr>
<td>Right</td>
<td>831 (23.2)</td>
</tr>
<tr>
<td>Unknown/subpleural</td>
<td>42 (1.1)</td>
</tr>
<tr>
<td><strong>Potential etiology</strong></td>
<td></td>
</tr>
<tr>
<td>Infectious/inflammatory</td>
<td>795 (31.8)</td>
</tr>
<tr>
<td>Malignant neoplasm</td>
<td>319 (12.8)</td>
</tr>
<tr>
<td>Other</td>
<td>291 (11.6)</td>
</tr>
<tr>
<td>More than 1 cause</td>
<td>1095 (43.8)</td>
</tr>
<tr>
<td><strong>Size</strong></td>
<td></td>
</tr>
<tr>
<td>&lt;6 mm</td>
<td>274 (11.6)</td>
</tr>
<tr>
<td>6-20 mm</td>
<td>1139 (48.5)</td>
</tr>
<tr>
<td>&gt;20 mm</td>
<td>340 (14.5)</td>
</tr>
<tr>
<td>More than 1 size</td>
<td>597 (25.4)</td>
</tr>
<tr>
<td><strong>GGO status</strong></td>
<td></td>
</tr>
<tr>
<td>Better</td>
<td>97 (4.2)</td>
</tr>
<tr>
<td>Stable</td>
<td>1388 (59.4)</td>
</tr>
<tr>
<td>Worse</td>
<td>288 (12.3)</td>
</tr>
<tr>
<td>More than 1 status</td>
<td>564 (24.1)</td>
</tr>
<tr>
<td><strong>Shape/margin</strong></td>
<td></td>
</tr>
<tr>
<td>Irregular/spiculated</td>
<td>845 (81)</td>
</tr>
<tr>
<td>Rounded/smooth</td>
<td>63 (6)</td>
</tr>
<tr>
<td>More than 1 shape</td>
<td>135 (13)</td>
</tr>
<tr>
<td><strong>Change in GGO size</strong></td>
<td></td>
</tr>
<tr>
<td>Increase in size</td>
<td>240 (16.8)</td>
</tr>
<tr>
<td>Decrease in size</td>
<td>208 (14.6)</td>
</tr>
<tr>
<td>Stable in size</td>
<td>976 (68.5)</td>
</tr>
<tr>
<td><strong>Change in GGO status</strong></td>
<td></td>
</tr>
<tr>
<td>Increased</td>
<td>259 (23)</td>
</tr>
<tr>
<td>Decreased</td>
<td>27 (2.4)</td>
</tr>
<tr>
<td>Stayed stable</td>
<td>815 (72.3)</td>
</tr>
<tr>
<td>Resolved</td>
<td>26 (2.3)</td>
</tr>
</tbody>
</table>

\(^a\)NLP: natural language processing.
\(^b\)GGO: ground-glass opacity.
\(^c\)Patient numbers were calculated from the first notes. GGO status was based on the description in the notes.
\(^d\)Longitudinal analysis between the first and the last notes.
\(^e\)Longitudinal 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>
<thead>
<tr>
<th>Size/timeline</th>
<th>&lt;6 months, n (%)</th>
<th>6 months to 1 year, n (%)</th>
<th>1 year to 3 years, n (%)</th>
<th>&gt;3 years, n (%)</th>
<th>Total, n (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>&lt;6 mm</td>
<td>61 (47.3)</td>
<td>17 (13.2)</td>
<td>29 (22.5)</td>
<td>22 (17)</td>
<td>129 (100)</td>
</tr>
<tr>
<td>6-20 mm</td>
<td>319 (58)</td>
<td>58 (10.5)</td>
<td>94 (17.1)</td>
<td>79 (14.4)</td>
<td>550 (100)</td>
</tr>
<tr>
<td>&gt;20 mm</td>
<td>136 (78.2)</td>
<td>9 (5.2)</td>
<td>14 (8)</td>
<td>15 (8.6)</td>
<td>174 (100)</td>
</tr>
</tbody>
</table>

Table 4. Patients in each size category at the different timelines from the first ground-glass opacity (GGO) notes to lung cancer diagnosis.
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 Sheikhshahi 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.

https://ai.jmir.org/2023/1/e44537

Lee et al
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, QF, 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, QF, and WO 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.

References


https://ai.jmir.org/2023/1/e44537

JMIR AI 2023 | vol. 2 | e44537 | p.261
(page number not for citation purposes)


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

©Kyeryoung Lee, Zongzhi Liu, Urmila Chandran, Iftekhar 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.
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

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. Ordoñez 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, 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.
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.

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.

References


27. MotionSense dataset. GitHub. URL: https://github.com/mmalekzadeh/motion-sense [accessed 2023-05-18]

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 Wright⁵, PhD; Sungjoon An⁶; Elaine Nguyen⁷, 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

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
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, biomedical 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 (CDECS) 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).

<table>
<thead>
<tr>
<th>Characteristics</th>
<th>Values</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Age category, n (%)</strong></td>
<td></td>
</tr>
<tr>
<td>Adults (aged 20-89 years)</td>
<td>38 (93)</td>
</tr>
<tr>
<td>Teenagers (aged 16-19 years)</td>
<td>3 (7)</td>
</tr>
<tr>
<td><strong>Age, (years), mean (SD)</strong></td>
<td>48.4 (20.4)</td>
</tr>
<tr>
<td><strong>Age, (years), median (IQR)</strong></td>
<td>48 (32-66)</td>
</tr>
<tr>
<td><strong>Gender a, n (%)</strong></td>
<td></td>
</tr>
<tr>
<td>Male</td>
<td>19 (46)</td>
</tr>
<tr>
<td>Female</td>
<td>21 (51)</td>
</tr>
<tr>
<td><strong>Diabetes type, n (%)</strong></td>
<td></td>
</tr>
<tr>
<td>Type 1</td>
<td>17 (41)</td>
</tr>
<tr>
<td>Type 2</td>
<td>24 (59)</td>
</tr>
<tr>
<td><strong>Race b, n (%)</strong></td>
<td></td>
</tr>
<tr>
<td>Alaska Native/American Indian</td>
<td>7 (17)</td>
</tr>
<tr>
<td>Black</td>
<td>13 (32)</td>
</tr>
<tr>
<td>White</td>
<td>24 (59)</td>
</tr>
<tr>
<td><strong>Advanced technology user a, n (%)</strong></td>
<td></td>
</tr>
<tr>
<td>Yes</td>
<td>33 (80)</td>
</tr>
<tr>
<td>No</td>
<td>7 (17)</td>
</tr>
<tr>
<td><strong>Education level, n (%)</strong></td>
<td></td>
</tr>
<tr>
<td>Some high school</td>
<td>3 (7)</td>
</tr>
<tr>
<td>High school, General Educational Development test, or equivalent</td>
<td>6 (15)</td>
</tr>
<tr>
<td>Trade school, apprenticeship, or equivalent</td>
<td>4 (10)</td>
</tr>
<tr>
<td>Associate’s degree</td>
<td>8 (20)</td>
</tr>
<tr>
<td>Bachelor’s degree</td>
<td>9 (22)</td>
</tr>
<tr>
<td>Postgraduate or professional degree</td>
<td>11 (27)</td>
</tr>
</tbody>
</table>

a Data for 40 participants only; percentages are of 41 participants and do not add up to 100.
b Not 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 patient 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, 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 the 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
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]


17. Barriers to Telehealth in Rural Areas. Rural Health Information Hub. URL: https://www.ruralhealthinfo.org/toolkits/telehealth/1/barriers [accessed 2023-05-19]


Applying Human Factors and Usability Engineering to Medical Devices. Food and Drug Administration. URL: https://tinyurl.com/24ae6are [accessed 2023-05-19]

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]


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:
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
PMID:
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 validation, 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

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 counties, 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.

<table>
<thead>
<tr>
<th>Criterion</th>
<th>Inclusion</th>
<th>Exclusion</th>
</tr>
</thead>
<tbody>
<tr>
<td>Type of study</td>
<td>Qualitative, quantitative, mixed method, and review studies in peer-reviewed journals</td>
<td>Letters, comments, conference abstracts, editorials, and theses</td>
</tr>
<tr>
<td>Language</td>
<td>English</td>
<td>All other languages</td>
</tr>
<tr>
<td>Study variables</td>
<td>Includes (1) artificial intelligence/machine learning and relevant terms and (2) allocation of health care resources</td>
<td>Does not include (1) artificial intelligence/machine learning and relevant terms or (2) allocation of health care resources</td>
</tr>
<tr>
<td>Study context</td>
<td>Health care resource allocation at either the population level or hospital level</td>
<td>All other resource allocation scenarios</td>
</tr>
</tbody>
</table>

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.
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.

<table>
<thead>
<tr>
<th>Reference</th>
<th>Objectives</th>
<th>Resource allocation situation</th>
<th>Input variables</th>
</tr>
</thead>
<tbody>
<tr>
<td>Rosas et al (2013) [25]</td>
<td>To construct a financial resource allocation model using an artificial neural network</td>
<td>Financial resources for public health in Brazil</td>
<td>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</td>
</tr>
<tr>
<td>Belciug &amp; Gorunescu (2015) [22]</td>
<td>To propose a bed allocation and financial resource utilization strategy through queuing modeling and evolutionary computation</td>
<td>Bed allocation and financial resource utilization in the geriatric department of a London hospital</td>
<td>Bed inventory, arrival rate, mean service time, patient flow parameters, and holding and penalty cost and other cost considerations</td>
</tr>
<tr>
<td>Gartner &amp; Padman (2015) [23]</td>
<td>To evaluate how early determination of diagnosis-related groups can be used for better allocation of scarce hospital resources</td>
<td>Hospital resources, including day resources and overnight resources (beds), validated in a mid-sized hospital near Munich, Germany</td>
<td>Primary and secondary diagnoses, clinical procedures, age, gender, and weight in newborns</td>
</tr>
<tr>
<td>Velez et al (2016) [26]</td>
<td>To present an artificial intelligence–based health planning model based on data from geospatial systems</td>
<td>Health care resource distribution in health planning in Ecuador</td>
<td>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</td>
</tr>
<tr>
<td>Xu et al (2018) [27]</td>
<td>To propose a health resource allocation model based on mass customization to maximize revenue and customization</td>
<td>Allocation of doctors and other medical resources in a public hospital system in China</td>
<td>Distribution of medical stations, professional level of doctors (salary and seniority), patient preferences and illness severity, medical cost, and revenue</td>
</tr>
<tr>
<td>Yousefi et al (2018) [28]</td>
<td>To present a model based on agent-based simulation, machine learning, and a genetic algorithm for allocation of medical resources in emergency departments</td>
<td>Medical resource allocation in a teaching hospital in the capital of State of Minas Gerais in Brazil</td>
<td>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 sputtering yellow zone, orthopedics department, surgical department, and clinical emergency area.</td>
</tr>
<tr>
<td>Zhang et al (2018) [29]</td>
<td>To propose a framework introducing a novel approach to multi-attribute decision-making problems in the picture fuzzy context</td>
<td>Medical resource allocation in the hierarchical medical treatment system in China</td>
<td>Patient diagnostic characteristics and hospital tiers</td>
</tr>
<tr>
<td>McRae et al (2020) [24]</td>
<td>To present a clinical decision-support system and mobile app to assist in COVID severity assessment, management, and care</td>
<td>Resource allocation during COVID in New York, with validation data sets from Wuhan, China</td>
<td>Outpatient score (age, gender, diabetes, cardiovascular comorbidities, and systolic blood pressure) and biomarker score (C-reactive protein, procalcitonin, and age)</td>
</tr>
<tr>
<td>Bednarski et al (2021) [21]</td>
<td>To study how reinforcement learning and deep-learning models can facilitate the redistribution of medical equipment during pandemics</td>
<td>Pandemics in the context of COVID</td>
<td>COVID risk factors by region, COVID mortality by region, and current demand for medical equipment</td>
</tr>
</tbody>
</table>

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.

<table>
<thead>
<tr>
<th>Reference</th>
<th>Objective</th>
<th>Resource allocation situation</th>
<th>Reviewed/discussed methods for the application of AI during the COVID-19 pandemic</th>
</tr>
</thead>
<tbody>
<tr>
<td>Rajkumar et al (2018) [40]</td>
<td>To explore how model design, biases in data, and interactions of model predictions with clinicians and patients exacerbate health inequalities</td>
<td>Intensive care unit and in-hospital stay length</td>
<td>• Suggested that future AI models for health care resource distribution should include principles of distributive justice.</td>
</tr>
<tr>
<td>Laudanski et al (2020) [36]</td>
<td>To analyze the applications of AI during COVID using the WHO framework of pandemic evolution</td>
<td>Global COVID-19 pandemic</td>
<td>• 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.</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>• Discussed AI-driven triage based on patient characteristics and AI-supported health resource allocation and ethics.</td>
</tr>
<tr>
<td>Adly et al (2020) [31]</td>
<td>To discuss the potential of AI to prevent and control COVID</td>
<td>Global COVID-19 pandemic</td>
<td>• Suggested that the application of AI was valuable in medical resource distribution that included the parameters of patients and the pandemic.</td>
</tr>
<tr>
<td>Bernardo et al (2020) [33]</td>
<td>To present approaches for using technology to facilitate resource distribution in disasters and outbreaks</td>
<td>Disasters and disease outbreaks</td>
<td>• Found that data collected from crowdsourcing and the human-technology interface could be used as data sources.</td>
</tr>
<tr>
<td>Neves et al (2020) [38]</td>
<td>To discuss the basic principles of medical resource allocation choices during COVID</td>
<td>Global COVID-19 pandemic</td>
<td>• Discussed rationalization of care, medical and team conflict, modeling of the pandemic, and application of AI.</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>• Explored the use of AI as a support tool to streamline inventory control and standardize resource distribution.</td>
</tr>
<tr>
<td>Xie et al (2020) [42]</td>
<td>To present an overview of the application of AI technology in ophthalmology, with a focus on deep-learning systems</td>
<td>Diabetic retinopathy screening</td>
<td>• Reviewed empirical considerations behind the formation of successful screening programs.</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>• Examined potential methods for health economics and safety analyses that can assess concerns regarding AI-based screening.</td>
</tr>
<tr>
<td>Zou et al (2020) [43]</td>
<td>To present the COVID response of Shenzhen, China and discuss the potential of a successful model for COVID prevention and control</td>
<td>COVID-19 pandemic in Shenzhen, China</td>
<td>• Reviewed methods applied by Shenzhen, including early action and centralized response, care for vulnerable persons, community response teams, and technology.</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>• Discussed the integration of information technology in Shenzhen’s response, including mobile technology, big data, and AI.</td>
</tr>
<tr>
<td>Basit et al (2021) [32]</td>
<td>To discuss the data sharing and collection process and the ethical considerations around pandemic data</td>
<td>Global COVID-19 pandemic</td>
<td>• 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.</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>• Discussed the application of AI in the allocation of intensive care resources and ventilators.</td>
</tr>
<tr>
<td>Huang et al (2021) [34]</td>
<td>To investigate China’s health informatization, especially during the COVID-19 pandemic</td>
<td>COVID-19 pandemic in China</td>
<td>• 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.</td>
</tr>
<tr>
<td>Jain et al (2021) [35]</td>
<td>To discuss the implications of AI for employability by analyzing issues in the health care sector</td>
<td>Human resources in health systems</td>
<td>• Displayed hierarchical relationships between employability and a range of characteristics.</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>• Discussed measures that could potentially enhance employability in the health care sector through AI.</td>
</tr>
<tr>
<td>Lu et al (2021) [37]</td>
<td>To establish barriers that affect medical information digitalization innovation and development through interviews and a literature review</td>
<td>Medical information digitalization</td>
<td>• Applied the importance-resistance analysis model and identified the resistant factors, including data sharing, infrastructure, regulation, and operations in the context of data privacy.</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>• Proposed several ways to overcome these limitations, including transparency regulation and infrastructure building.</td>
</tr>
</tbody>
</table>
Reviewed/discussed methods for the application of AI during the COVID-19 pandemic

- 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.
- Discussed how COVID exacerbated racial and socioeconomic disparities.
- Explored how an AI-informed resource allocation strategy can be influenced by biases.

Objective

- To present interindividual variability and the roles it plays in the variability of COVID presentation and susceptibility.
- To discuss possible bias in the application of AI during the COVID-19 pandemic.

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 [21,22,23-34,37,39,42,43] 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.

<table>
<thead>
<tr>
<th>Reference</th>
<th>Policy relevance</th>
</tr>
</thead>
</table>
<pre><code>                             | - 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. |
</code></pre>
| Belciug & Gorunescu (2015) [22]  | - Provided tools to estimate the appropriate parameters for optimal resource utilization.  
                                 | - Enabled the hospital manager to simulate scenarios to make the near-best decision.                                                                                                                         |
                                 | - Offered an approach to integrate and analyze the financial objectives of health care delivery.                                                                                                           |
| Velez et al (2016) [26]          | - 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]             | - Reduced costs by making doctors mobile.  
                                 | - Addressed personal preferences, such as treatment time and the professional level of doctors.                                                                                                               |
| Yousefi et al (2018) [28]        | - 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]          | - 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]          | - 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.                                                                                                                       |
                                 | - Improved algorithm performance for future applications.                                                                                                                                                |
They argued that the choice of input variables for health care mortality, socioeconomic characteristics, and income inequality. algorithm for the public hospital system in Brazil based on is worth exploring.
mortalities, the resource distribution situation in the 2 countries development and the enormous difference in COVID cases and 
Given the similarity between the 2 countries in economic 
total of 124,000 cases and 4636 deaths as of March 2022 [45]. China 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].

Table 5. Policy relevance of the included studies involving theoretical discussions, qualitative studies, or review studies.

<table>
<thead>
<tr>
<th>Reference</th>
<th>Policy relevance</th>
</tr>
</thead>
<tbody>
<tr>
<td>Rajkomar et al (2018) [40]</td>
<td>Proposed that the principles of distributive justice be incorporated into model design, deployment, and evaluation.</td>
</tr>
<tr>
<td>Laudanski et al (2020) [36]</td>
<td>Suggested that AI can couple outbreak data with measures of potential demand and direct supplies more efficiently.</td>
</tr>
<tr>
<td>Adly et al (2020) [31]</td>
<td>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.</td>
</tr>
<tr>
<td>Bernardo et al (2020) [33]</td>
<td>Suggested that automation by AI and machine learning can further our abilities in predictive analytics.</td>
</tr>
<tr>
<td>Neves et al (2020) [38]</td>
<td>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.</td>
</tr>
<tr>
<td>Xie et al (2020) [42]</td>
<td>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.</td>
</tr>
<tr>
<td>Zou et al (2020) [43]</td>
<td>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.</td>
</tr>
<tr>
<td>Basit et al (2021) [32]</td>
<td>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.</td>
</tr>
<tr>
<td>Huang et al (2021) [34]</td>
<td>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.</td>
</tr>
<tr>
<td>Jain et al (2021) [35]</td>
<td>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.</td>
</tr>
<tr>
<td>Lu et al (2021) [37]</td>
<td>Provided a basis for the future development directions of medical information digitalization and its impacts on health care and health systems.</td>
</tr>
<tr>
<td>Pereira et al (2021) [39]</td>
<td>Suggested that predicting which COVID-19 patients will develop progressive diseases that require hospitalization has important implications for clinical trials targeting outpatients.</td>
</tr>
<tr>
<td>Roöstl et al (2021) [41]</td>
<td>Proposed that transparency in reporting of AI algorithms is necessary to understand intended predictions, target populations, hidden biases, and class imbalance problems.</td>
</tr>
</tbody>
</table>

*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].

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

Abbreviations

AI: artificial intelligence
Predicting Treatment Interruption Among People Living With HIV in Nigeria: Machine Learning Approach

Matthew-David Ogbechie¹*, BSc; Christa Fischer Walker²*, 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 Persaud²*, 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

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
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.

<table>
<thead>
<tr>
<th>Variable and category</th>
<th>Individuals (N=136,747), n (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Interruption in treatment</strong></td>
<td></td>
</tr>
<tr>
<td>Yes</td>
<td>56,581 (41.38)</td>
</tr>
<tr>
<td>No</td>
<td>80,166 (58.62)</td>
</tr>
<tr>
<td><strong>Year of antiretroviral initiation</strong></td>
<td></td>
</tr>
<tr>
<td>Before 2017</td>
<td>61,939 (45.3)</td>
</tr>
<tr>
<td>January 2017-September 2019</td>
<td>46,776 (34.2)</td>
</tr>
<tr>
<td>October 2019-February 2021</td>
<td>28,032 (20.5)</td>
</tr>
<tr>
<td><strong>Gender</strong></td>
<td></td>
</tr>
<tr>
<td>Female</td>
<td>91,982 (67.26)</td>
</tr>
<tr>
<td>Male</td>
<td>44,765 (32.74)</td>
</tr>
<tr>
<td><strong>Age (years)</strong></td>
<td></td>
</tr>
<tr>
<td>&lt;14</td>
<td>5657 (4.14)</td>
</tr>
<tr>
<td>14-20</td>
<td>8685 (6.35)</td>
</tr>
<tr>
<td>21-35</td>
<td>72,049 (52.69)</td>
</tr>
<tr>
<td>&gt;35</td>
<td>50,356 (36.82)</td>
</tr>
<tr>
<td><strong>Marital status</strong></td>
<td></td>
</tr>
<tr>
<td>Married</td>
<td>1171 (0.86)</td>
</tr>
<tr>
<td>Single</td>
<td>64,899 (47.46)</td>
</tr>
<tr>
<td>Previously married</td>
<td>52,934 (38.71)</td>
</tr>
<tr>
<td><strong>Education</strong></td>
<td></td>
</tr>
<tr>
<td>Primary and Quranic</td>
<td>16,473 (12.05)</td>
</tr>
<tr>
<td>≥1 year of secondary</td>
<td>44,219 (32.34)</td>
</tr>
<tr>
<td>None</td>
<td>50,912 (37.23)</td>
</tr>
<tr>
<td><strong>Occupation</strong></td>
<td></td>
</tr>
<tr>
<td>Employed</td>
<td>36,863 (26.96)</td>
</tr>
<tr>
<td>Unemployed/retired/students</td>
<td>79,059 (57.81)</td>
</tr>
<tr>
<td><strong>State</strong></td>
<td></td>
</tr>
<tr>
<td>Akwa Ibom</td>
<td>100,937 (73.82)</td>
</tr>
<tr>
<td>Cross River</td>
<td>35,791 (26.18)</td>
</tr>
<tr>
<td><strong>Baseline clinic stage</strong></td>
<td></td>
</tr>
<tr>
<td>Stage I</td>
<td>64,508 (47.17)</td>
</tr>
<tr>
<td>Stage II-IV</td>
<td>68,740 (50.27)</td>
</tr>
<tr>
<td><strong>Facility type</strong></td>
<td></td>
</tr>
<tr>
<td>Health center/clinic/posts</td>
<td>77,597 (56.75)</td>
</tr>
<tr>
<td>General, tertiary, or cottage hospital</td>
<td>59,131 (43.25)</td>
</tr>
<tr>
<td><strong>TB* status</strong></td>
<td></td>
</tr>
<tr>
<td>No signs or symptoms of TB</td>
<td>58,953 (43.11)</td>
</tr>
<tr>
<td>Currently on isoniazid prophylaxis</td>
<td>4745 (3.47)</td>
</tr>
<tr>
<td>Confirmed/suspected TB</td>
<td>5167 (3.8)</td>
</tr>
<tr>
<td><strong>Pregnant</strong></td>
<td></td>
</tr>
<tr>
<td>No</td>
<td>46,883 (51.0)</td>
</tr>
</tbody>
</table>
Table 2. Model performance evaluation with test data from January 2005 through February 2021 and validation data for model 4.

<table>
<thead>
<tr>
<th>Variable and category</th>
<th>Individuals (N=136,747), n (%)</th>
<th>Model 1</th>
<th>Model 2</th>
<th>Model 3</th>
<th>Model 4 (selected)</th>
<th>Model 5</th>
<th>Model 6</th>
<th>Model 7</th>
<th>Model 8</th>
</tr>
</thead>
<tbody>
<tr>
<td>Yes</td>
<td>925 (1.0)</td>
<td>0.83 (0.85-0.86)</td>
<td>0.83 (0.83-0.84)</td>
<td>0.87 (0.87-0.87)</td>
<td>0.85 (0.85-0.86)</td>
<td>0.91 (0.88-0.93)</td>
<td>0.75 (0.74-0.75)</td>
<td>0.70 (0.69-0.70)</td>
<td>0.75 (0.74-0.75)</td>
</tr>
<tr>
<td>Breastfeeding&lt;sup&gt;c&lt;/sup&gt;</td>
<td>No</td>
<td>47,617 (51.8)</td>
<td>0.82 (0.81-0.82)</td>
<td>0.75 (0.75-0.76)</td>
<td>0.84 (0.83-0.84)</td>
<td>0.81 (0.81-0.82)</td>
<td>0.79 (0.73-0.86)</td>
<td>0.63 (0.62-0.64)</td>
<td>0.58 (0.57-0.59)</td>
</tr>
<tr>
<td></td>
<td>Yes</td>
<td>191 (0.21)</td>
<td>0.88 (0.87-0.88)</td>
<td>0.89 (0.88-0.89)</td>
<td>0.89 (0.89-0.90)</td>
<td>0.88 (0.88-0.89)</td>
<td>0.94 (0.92-0.96)</td>
<td>0.83 (0.82-0.84)</td>
<td>0.78 (0.77-0.78)</td>
</tr>
</tbody>
</table>

<sup>a</sup>TB: tuberculosis.
<sup>b</sup>Totals do not add up to 136,747 for all variables under TB status due to missing values for some variables.
<sup>c</sup>n=91,982 (number of females in the data set).

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:

https://ai.jmir.org/2023/1/e44432
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 indicators 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 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


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

©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

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 (HbA1c) 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 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].

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 HbA1c 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 HbA1c 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 HbA1c measurement fails to effectively capture these oscillations, HbA1c level alone is not an ideal indicator of an individual patient’s glycemic control [14].

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 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 predicted values. Elij 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).

<table>
<thead>
<tr>
<th>Metrics and analytic methods</th>
<th>Related publications</th>
<th>Strengths</th>
<th>Limitations</th>
</tr>
</thead>
</table>
| SD                           | Krinsley [19]        | Simplicity | • 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 |
| COV<sup>b</sup>             | Rodbard [37] and Rama Chandran et al [36] | Simplicity and adjusts for mean | • Does not account for temporal information  
• Limited capability of interpreting GV profiles with only a single value |
| TIR<sup>c</sup>             | 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 | • Does not account for sequence of BG measurements |
| IQR                          | McDonnel et al [39]  | Simplicity | • 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 | • 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<sup>d</sup>            | Service [22] and Service et al [40] | Takes BG fluctuation owing to meal into account | • 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<sup>e</sup> and HBGI<sup>f</sup> | Kovatchev et al [41] and Hill et al [42] | Adjusts for BG skewness and measuring frequency | • 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<sup>g</sup>             | Struble [31] and Marling et al [32] | Accounts for temporal information | • 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<sup>h</sup>          | Eljil et al [33]     | Accounts for temporal information | • Limited capability of interpreting GV profiles with only a single value |
| RF<sup>i</sup>              | Mani et al [34]      | Accounts for temporal information | • Limited capability of interpreting GV profiles with only a single value |
| Glucotypes                  | Hall et al [35]      | Accounts for temporal information | • Limited capability of interpreting GV profiles with only 3 discrete levels |

<sup>a</sup>BG: blood glucose.  
<sup>b</sup>COV: coefficient of variation.  
<sup>c</sup>TIR: time in range.  
<sup>d</sup>MAGE: mean amplitude of glycemic excursions.  
<sup>e</sup>LBGI: low blood glucose index.  
<sup>f</sup>HBGI: high blood glucose index.  
<sup>g</sup>SVR: support vector regression.  
<sup>h</sup>TS-ANN: time-sensitive artificial neural network.  
<sup>i</sup>RF: 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\textsubscript{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.
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 HbA1c 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 HbA1c 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.

<table>
<thead>
<tr>
<th>GV pattern number</th>
<th>Glucose level</th>
<th>Pattern trends</th>
<th>Occurrence (N=149,639), n (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>6</td>
<td>Severely hyperglycemic</td>
<td>Steady or rising to peak and declining</td>
<td>8440 (5.64)</td>
</tr>
<tr>
<td>5</td>
<td>Hyperglycemic</td>
<td>Steady or concave up or down</td>
<td>22,594 (15.10)</td>
</tr>
<tr>
<td>4</td>
<td>From marginally hyperglycemic to hyperglycemic</td>
<td>Rising</td>
<td>28,653 (19.15)</td>
</tr>
<tr>
<td>3</td>
<td>From marginally hyperglycemic to normal</td>
<td>Declining</td>
<td>31,185 (20.84)</td>
</tr>
<tr>
<td>2</td>
<td>Normal</td>
<td>Steady or slightly rising</td>
<td>30,255 (20.22)</td>
</tr>
<tr>
<td>1</td>
<td>Marginally hypoglycemic or normal</td>
<td>Steady or concave up</td>
<td>28,512 (19.05)</td>
</tr>
</tbody>
</table>

**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.

https://ai.jmir.org/2023/1/e45450

 JMIR AI 2023 | vol. 2 | e45450 | p.316
 (page number not for citation purposes)
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).

<table>
<thead>
<tr>
<th>Characteristics</th>
<th>Cluster A (n=74)</th>
<th>Cluster B (n=4)</th>
<th>Cluster C (n=35)</th>
<th>Cluster D (n=13)</th>
<th>P value&lt;sup&gt;a&lt;/sup&gt;</th>
</tr>
</thead>
<tbody>
<tr>
<td>Age (years), mean (SD)</td>
<td>40.3 (14.3)</td>
<td>22.8 (4.27)</td>
<td>41.8 (12.9)</td>
<td>33.8 (10.7)</td>
<td>.02</td>
</tr>
<tr>
<td>Sex (female), n (%)</td>
<td>34 (46)</td>
<td>1 (25)</td>
<td>18 (51)</td>
<td>9 (69)</td>
<td>.33</td>
</tr>
<tr>
<td>Index of Multiple Deprivation decile [53], mean (SD)</td>
<td>7.96 (2.31)</td>
<td>8.25 (2.06)</td>
<td>7.34 (2.44)</td>
<td>7.38 (2.66)</td>
<td>.56</td>
</tr>
<tr>
<td>BMI (kg/m²), mean (SD)</td>
<td>27.3 (4.64)</td>
<td>24.1 (3.07)</td>
<td>26.9 (5.69)</td>
<td>23.4 (4.39)</td>
<td>.22</td>
</tr>
<tr>
<td>Duration of diabetes (years), mean (SD)</td>
<td>22.2 (12.4)</td>
<td>8.5 (4.36)</td>
<td>24.8 (15.7)</td>
<td>14.5 (14.7)</td>
<td>.04</td>
</tr>
<tr>
<td>Number of days since CGM&lt;sup&gt;b&lt;/sup&gt; use, mean (SD)</td>
<td>218 (243)</td>
<td>203 (56.1)</td>
<td>167 (203)</td>
<td>215 (276)</td>
<td>.76</td>
</tr>
<tr>
<td>eGFR&lt;sup&gt;c&lt;/sup&gt; stage, n (%)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>.85</td>
</tr>
<tr>
<td>5</td>
<td>0 (0)</td>
<td>0 (0)</td>
<td>0 (0)</td>
<td>0 (0)</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>1 (1)</td>
<td>0 (0)</td>
<td>0 (0)</td>
<td>0 (0)</td>
<td></td>
</tr>
<tr>
<td>3b</td>
<td>2 (3)</td>
<td>0 (0)</td>
<td>0 (0)</td>
<td>0 (0)</td>
<td></td>
</tr>
<tr>
<td>3a</td>
<td>4 (5)</td>
<td>0 (0)</td>
<td>1 (3)</td>
<td>0 (0)</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>28 (38)</td>
<td>0 (0)</td>
<td>16 (46)</td>
<td>6 (46)</td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>38 (51)</td>
<td>4 (100)</td>
<td>18 (51)</td>
<td>7 (54)</td>
<td></td>
</tr>
<tr>
<td>HbA&lt;sub&gt;1c&lt;/sub&gt;&lt;sup&gt;d&lt;/sup&gt; level (mmol/mol), n (%)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>&lt;.001</td>
</tr>
<tr>
<td>≤42</td>
<td>1 (1)</td>
<td>0 (0)</td>
<td>1 (3)</td>
<td>7 (54)</td>
<td></td>
</tr>
<tr>
<td>43-48</td>
<td>4 (5)</td>
<td>0 (0)</td>
<td>6 (17)</td>
<td>3 (23)</td>
<td></td>
</tr>
<tr>
<td>48-59</td>
<td>14 (19)</td>
<td>0 (0)</td>
<td>19 (54)</td>
<td>2 (15)</td>
<td></td>
</tr>
<tr>
<td>59-85</td>
<td>46 (62)</td>
<td>1 (25)</td>
<td>9 (26)</td>
<td>1 (8)</td>
<td></td>
</tr>
<tr>
<td>≥86</td>
<td>8 (11)</td>
<td>3 (75)</td>
<td>0 (0)</td>
<td>0 (0)</td>
<td></td>
</tr>
<tr>
<td>Glucose level, mean (SD)</td>
<td>10.8 (1.57)</td>
<td>19.3 (1.46)</td>
<td>8.22 (0.614)</td>
<td>6.48 (1.1)</td>
<td>&lt;.001</td>
</tr>
<tr>
<td>COV&lt;sup&gt;e&lt;/sup&gt; of glucose level, mean (SD)</td>
<td>0.428 (0.066)</td>
<td>0.354 (0.031)</td>
<td>0.429 (0.061)</td>
<td>0.37 (0.063)</td>
<td>.003</td>
</tr>
<tr>
<td>TIR&lt;sup&gt;f&lt;/sup&gt; (mmol/L), mean % (SD)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>≤3</td>
<td>1.9 (2.2)</td>
<td>0.3 (0.2)</td>
<td>3.2 (2.8)</td>
<td>4.6 (5)</td>
<td>.002</td>
</tr>
<tr>
<td>3-3.9</td>
<td>3.5 (2)</td>
<td>0.6 (0.2)</td>
<td>6.7 (2.7)</td>
<td>11 (7.7)</td>
<td>&lt;.001</td>
</tr>
<tr>
<td>3.9-10</td>
<td>43.3 (11.3)</td>
<td>10.6 (2.9)</td>
<td>62.4 (6.4)</td>
<td>74.7 (12.9)</td>
<td>&lt;.001</td>
</tr>
<tr>
<td>10-13.9</td>
<td>27.2 (6.0)</td>
<td>13.3 (3.6)</td>
<td>20.5 (4)</td>
<td>7.5 (4.9)</td>
<td>&lt;.001</td>
</tr>
<tr>
<td>≥13.9</td>
<td>24.1 (11.6)</td>
<td>75.2 (5.8)</td>
<td>7.3 (3.4)</td>
<td>2.3 (5.8)</td>
<td>&lt;.001</td>
</tr>
<tr>
<td>Time in patterns, mean % (SD)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>13.6 (7.2)</td>
<td>2.1 (0.6)</td>
<td>27.6 (8.8)</td>
<td>51.9 (17.9)</td>
<td>&lt;.001</td>
</tr>
<tr>
<td>2</td>
<td>17 (6.1)</td>
<td>3.6 (1.1)</td>
<td>27.6 (5)</td>
<td>32.5 (10.6)</td>
<td>&lt;.001</td>
</tr>
<tr>
<td>3</td>
<td>22.4 (4.7)</td>
<td>6.8 (1.7)</td>
<td>23.5 (5.2)</td>
<td>10.2 (6.6)</td>
<td>&lt;.001</td>
</tr>
<tr>
<td>4</td>
<td>23.1 (5.7)</td>
<td>10.1 (2.8)</td>
<td>15.6 (4.2)</td>
<td>5 (7.3)</td>
<td>&lt;.001</td>
</tr>
<tr>
<td>5</td>
<td>19.3 (7.8)</td>
<td>26.2 (5.8)</td>
<td>5.5 (3)</td>
<td>0.4 (0.7)</td>
<td>&lt;.001</td>
</tr>
<tr>
<td>6</td>
<td>4.7 (6)</td>
<td>51.2 (11)</td>
<td>0.3 (0.4)</td>
<td>0 (0)</td>
<td>&lt;.001</td>
</tr>
<tr>
<td>Fulfillment of recommended targets [28], n (%)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>TIR between 3.9 and 10 mmol/L &gt;70% of the time</td>
<td>7 (10)</td>
<td>0 (0)</td>
<td>23 (66)</td>
<td>11 (85)</td>
<td>&lt;.001</td>
</tr>
<tr>
<td>COV of glucose level &lt;0.36</td>
<td>8 (11)</td>
<td>2 (50)</td>
<td>4 (11)</td>
<td>7 (54)</td>
<td>&lt;.001</td>
</tr>
<tr>
<td>HbA&lt;sub&gt;1c&lt;/sub&gt; level &lt;58 mmol/mol</td>
<td>18 (24)</td>
<td>0 (0)</td>
<td>26 (74)</td>
<td>12 (92)</td>
<td>&lt;.001</td>
</tr>
<tr>
<td>Comorbidities, n (%)</td>
<td>55 (74)</td>
<td>3 (75)</td>
<td>22 (63)</td>
<td>5 (39)</td>
<td>.10</td>
</tr>
</tbody>
</table>
**Table**

<table>
<thead>
<tr>
<th>Characteristics</th>
<th>Cluster A (n=74)</th>
<th>Cluster B (n=4)</th>
<th>Cluster C (n=35)</th>
<th>Cluster D (n=13)</th>
<th>P value^a</th>
</tr>
</thead>
<tbody>
<tr>
<td>Diabetic complications, n (%)</td>
<td>47 (64)</td>
<td>3 (75)</td>
<td>19 (54)</td>
<td>4 (31)</td>
<td>.18</td>
</tr>
<tr>
<td>Medication, n (%)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Insulin</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Injection</td>
<td>66 (89)</td>
<td>4 (100)</td>
<td>33 (94)</td>
<td>12 (92)</td>
<td>.75</td>
</tr>
<tr>
<td>Pump</td>
<td>11 (15)</td>
<td>0 (0)</td>
<td>4 (11)</td>
<td>2 (15)</td>
<td>.78</td>
</tr>
<tr>
<td>Blood pressure</td>
<td>19 (26)</td>
<td>1 (25)</td>
<td>9 (26)</td>
<td>1 (8)</td>
<td>.55</td>
</tr>
<tr>
<td>Cholesterol</td>
<td>19 (26)</td>
<td>0 (0)</td>
<td>8 (23)</td>
<td>1 (8)</td>
<td>.33</td>
</tr>
<tr>
<td>Thyroid</td>
<td>10 (14)</td>
<td>0 (0)</td>
<td>2 (6)</td>
<td>2 (15)</td>
<td>.50</td>
</tr>
<tr>
<td>Antiplatelet</td>
<td>4 (5)</td>
<td>0 (0)</td>
<td>2 (6)</td>
<td>1 (8)</td>
<td>.95</td>
</tr>
<tr>
<td>Psychology</td>
<td>7 (10)</td>
<td>0 (0)</td>
<td>0 (0)</td>
<td>0 (0)</td>
<td>.15</td>
</tr>
</tbody>
</table>

^aP values <.05 are italicized; missing values were omitted only during the calculation of P values.

**Characteristics and Medication**

- **Insulin**: Injection and Pump are compared among clusters.
- **Diabetic complications**: A new column is added for this category.
- **Cholesterol**, **Blood pressure**, **Thyroid**, **Antiplatelet**, and **Psychology** are also included.

**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 glycemic 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 HbA1c 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 glycemic targets (all \( P > .05 \); *Table 4*), long duration is associated with specific glycemic metrics, including high mean glucose (\( P = .03 \)), TIR \( \geq 13.9 \) 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).

<table>
<thead>
<tr>
<th>Characteristics</th>
<th>&lt;68 days (n=46)</th>
<th>68-180 days (n=40)</th>
<th>&gt;180 days (n=40)</th>
<th>P value&lt;sup&gt;a&lt;/sup&gt;</th>
</tr>
</thead>
<tbody>
<tr>
<td>Age (years), mean (SD)</td>
<td>38.7 (12.5)</td>
<td>40.4 (15.2)</td>
<td>39.4 (14)</td>
<td>.87</td>
</tr>
<tr>
<td>Sex (female), n (%)</td>
<td>26 (57)</td>
<td>16 (40)</td>
<td>20 (50)</td>
<td>.31</td>
</tr>
<tr>
<td>Index of Multiple Deprivation decile [53], mean (SD)</td>
<td>7.83 (2.01)</td>
<td>8.1 (2.34)</td>
<td>7.28 (2.73)</td>
<td>.28</td>
</tr>
<tr>
<td>BMI (kg/m&lt;sup&gt;2&lt;/sup&gt;), mean (SD)</td>
<td>27.8 (6.2)</td>
<td>26.8 (3.9)</td>
<td>25.7 (4)</td>
<td>.22</td>
</tr>
<tr>
<td>Glucose level, mean (SD)</td>
<td>9.08 (1.57)</td>
<td>10.2 (2.5)</td>
<td>10.5 (3.5)</td>
<td>.03</td>
</tr>
<tr>
<td>COV&lt;sup&gt;b&lt;/sup&gt; of glucose level, mean (SD)</td>
<td>0.434 (0.074)</td>
<td>0.416 (0.06)</td>
<td>0.408 (0.061)</td>
<td>.17</td>
</tr>
<tr>
<td><strong>TIR</strong>&lt;sup&gt;c&lt;/sup&gt; (mmol/L), mean % (SD)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>≤3</td>
<td>3 (3.5)</td>
<td>1.7 (1.5)</td>
<td>2.6 (3)</td>
<td>.11</td>
</tr>
<tr>
<td>3.9-10</td>
<td>5.9 (3.6)</td>
<td>4.3 (2.8)</td>
<td>5 (5.3)</td>
<td>.19</td>
</tr>
<tr>
<td>10-13.9</td>
<td>55.3 (13.5)</td>
<td>49.1 (17.3)</td>
<td>47.3 (18.9)</td>
<td>.07</td>
</tr>
<tr>
<td>≥13.9</td>
<td>21.7 (7.2)</td>
<td>24.3 (8.5)</td>
<td>22.8 (9)</td>
<td>.36</td>
</tr>
<tr>
<td><strong>Time in patterns, mean % (SD)</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>24.7 (14.5)</td>
<td>17.4 (13.5)</td>
<td>20.5 (17.5)</td>
<td>.09</td>
</tr>
<tr>
<td>2</td>
<td>23.4 (8.8)</td>
<td>20.2 (8.8)</td>
<td>19.4 (9.7)</td>
<td>.10</td>
</tr>
<tr>
<td>3</td>
<td>21.2 (7)</td>
<td>21.7 (6.3)</td>
<td>19.9 (7.1)</td>
<td>.46</td>
</tr>
<tr>
<td>4</td>
<td>17.8 (7.8)</td>
<td>20.1 (8.2)</td>
<td>18.3 (8.3)</td>
<td>.40</td>
</tr>
<tr>
<td>5</td>
<td>11.1 (9.2)</td>
<td>15.6 (9.4)</td>
<td>14.8 (11.1)</td>
<td>.08</td>
</tr>
<tr>
<td>6</td>
<td>1.7 (2.7)</td>
<td>4.9 (8.7)</td>
<td>7.1 (14.9)</td>
<td>.04</td>
</tr>
<tr>
<td><strong>Fulfillment of recommended targets [28], n (%)</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>TIR between 3.9 and 10 mmol/L &gt;70% of the time</td>
<td>17 (37)</td>
<td>11 (28)</td>
<td>13 (33)</td>
<td>.65</td>
</tr>
<tr>
<td>COV of glucose level &lt;0.36</td>
<td>7 (15)</td>
<td>8 (20)</td>
<td>6 (15)</td>
<td>.79</td>
</tr>
<tr>
<td>HbA&lt;sub&gt;1c&lt;/sub&gt;&lt;sup&gt;d&lt;/sup&gt; level &lt;58 mmol/mol</td>
<td>19 (41)</td>
<td>17 (43)</td>
<td>20 (50)</td>
<td>.61</td>
</tr>
</tbody>
</table>

<sup>a</sup>P values <.05 are italicized; missing values were omitted only during the calculation of P values.

<sup>b</sup>COV: coefficient of variation.

<sup>c</sup>TIR: time in range.

<sup>d</sup>HbA<sub>1c</sub>: 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≥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≤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


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
- HbA1c: glycated hemoglobin
- IMD: Index of Multiple Deprivation
- NHS: National Health Service
- T1DM: type 1 diabetes mellitus
- TIR: time in range

©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.
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: winstonrlia@gmail.com

Abstract

Background: Nearly one-third of patients with diabetes are poorly controlled (hemoglobin A₁c ≥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

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 A1c (HbA1c) 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 HbA1c (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” [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 HbA1c 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.

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 use the tool, perceived usefulness, ease of use, and organizational support for use. Next, they ranked 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 A1c > 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).

<table>
<thead>
<tr>
<th>Characteristics</th>
<th>Values, n (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Gender</strong></td>
<td></td>
</tr>
<tr>
<td>Women</td>
<td>13 (59)</td>
</tr>
<tr>
<td>Men</td>
<td>8 (36)</td>
</tr>
<tr>
<td>Prefer not to answer</td>
<td>1 (5)</td>
</tr>
<tr>
<td><strong>Race and ethnicity (select all that apply)</strong></td>
<td></td>
</tr>
<tr>
<td>Hispanic, Latinx, or Spanish origin</td>
<td>9 (41)</td>
</tr>
<tr>
<td>White</td>
<td>6 (27)</td>
</tr>
<tr>
<td>Asian</td>
<td>4 (18)</td>
</tr>
<tr>
<td>Black or African American</td>
<td>1 (5)</td>
</tr>
<tr>
<td>Middle Eastern or North African</td>
<td>1 (5)</td>
</tr>
<tr>
<td>Prefer not to answer</td>
<td>1 (5)</td>
</tr>
<tr>
<td><strong>Professional role</strong></td>
<td></td>
</tr>
<tr>
<td>Physician</td>
<td>14 (64)</td>
</tr>
<tr>
<td>Nurse practitioner</td>
<td>1 (5)</td>
</tr>
<tr>
<td>Physician assistant</td>
<td>1 (5)</td>
</tr>
<tr>
<td>Nurse</td>
<td>1 (5)</td>
</tr>
<tr>
<td>Behavioral specialist</td>
<td>1 (5)</td>
</tr>
<tr>
<td>Social worker</td>
<td>1 (5)</td>
</tr>
<tr>
<td>Other (front desk, administrative, or medical assistant)</td>
<td>3 (14)</td>
</tr>
<tr>
<td><strong>Primary practice site</strong></td>
<td></td>
</tr>
<tr>
<td>Academic health center or faculty practice</td>
<td>8 (36)</td>
</tr>
<tr>
<td>Federally qualified health center or look-alike</td>
<td>6 (27)</td>
</tr>
<tr>
<td>Private solo or group practice</td>
<td>4 (18)</td>
</tr>
<tr>
<td>Health maintenance organization (eg, Kaiser Permanente)</td>
<td>2 (9)</td>
</tr>
<tr>
<td>Mental health center</td>
<td>1 (5)</td>
</tr>
<tr>
<td>Other (multiple sites)</td>
<td>1 (5)</td>
</tr>
<tr>
<td><strong>Specialty (includes physicians, nurse practitioners, and physician assistants)</strong></td>
<td></td>
</tr>
<tr>
<td>Family medicine</td>
<td>15 (94)</td>
</tr>
<tr>
<td>Pediatrics</td>
<td>1 (6)</td>
</tr>
<tr>
<td><strong>Years since residency graduation (physicians only)</strong></td>
<td></td>
</tr>
<tr>
<td>In residency</td>
<td>4 (29)</td>
</tr>
<tr>
<td>1-10</td>
<td>4 (29)</td>
</tr>
<tr>
<td>11-20</td>
<td>3 (21)</td>
</tr>
<tr>
<td>21-30</td>
<td>3 (21)</td>
</tr>
</tbody>
</table>
Table 2. Attitudes toward the tool and factors affecting implementation.

<table>
<thead>
<tr>
<th>Attitudes^a, mean (SD)</th>
<th>Values</th>
</tr>
</thead>
<tbody>
<tr>
<td>“I would use the clinical decision support tool.”</td>
<td>5.6 (1.4)</td>
</tr>
<tr>
<td>“I find the clinical decision support tool to be useful in my job.”</td>
<td>5.7 (1.3)</td>
</tr>
<tr>
<td>“I find the clinical decision support tool to be easy to use.”</td>
<td>5.8 (1.2)</td>
</tr>
<tr>
<td>“In general, the clinic would support my use of this clinical decision support tool.”</td>
<td>5.0 (1.7)</td>
</tr>
</tbody>
</table>

Factors affecting implementation (rank order)^d

<table>
<thead>
<tr>
<th>Factor, mean (SD)</th>
<th>Values</th>
</tr>
</thead>
<tbody>
<tr>
<td>Whether its use improves health</td>
<td>2.5 (1.7)</td>
</tr>
<tr>
<td>Accuracy</td>
<td>2.7 (1.7)</td>
</tr>
<tr>
<td>Usability</td>
<td>3.7 (1.5)</td>
</tr>
<tr>
<td>Impact on clinic workflows</td>
<td>3.9 (1.6)</td>
</tr>
<tr>
<td>Cost</td>
<td>4.2 (1.7)</td>
</tr>
<tr>
<td>Whether its use reduces costs to the health care system</td>
<td>4.3 (1.6)</td>
</tr>
</tbody>
</table>

A team of clinicians and staff were tasked with predicting whether the 1000 individuals with diabetes in your practice would have a hemoglobin A1c >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

<table>
<thead>
<tr>
<th>Distribution of responses, n (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>0-200</td>
</tr>
<tr>
<td>201-400</td>
</tr>
<tr>
<td>401-600</td>
</tr>
<tr>
<td>601-800</td>
</tr>
<tr>
<td>801-1000</td>
</tr>
</tbody>
</table>

^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).

<table>
<thead>
<tr>
<th>Themes and subthemes</th>
<th>Participants, n (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>How could the tool affect the delivery of care?</strong></td>
<td></td>
</tr>
<tr>
<td>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</td>
<td></td>
</tr>
<tr>
<td>The tool is not currently available, addresses a clinical gap, and represents a departure from the status quo.</td>
<td>7 (32)</td>
</tr>
<tr>
<td>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.</td>
<td>20 (91)</td>
</tr>
<tr>
<td>This tool could improve population health, address quality measures, and contribute to efficient resource allocation.</td>
<td>10 (45)</td>
</tr>
<tr>
<td>The tool would facilitate individualized and holistic care, by integrating primary care, behavioral health, and social care.</td>
<td>11 (50)</td>
</tr>
<tr>
<td>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.</td>
<td>7 (32)</td>
</tr>
<tr>
<td><strong>What concerns do clinicians and staff have about the tool?</strong></td>
<td></td>
</tr>
<tr>
<td>Clinicians and staff were concerned about how the tool affects patient-oriented outcomes and clinic workflows</td>
<td></td>
</tr>
<tr>
<td>Participants were concerned the tool would lead to harms, be used punitively, and make care more expensive.</td>
<td>15 (68)</td>
</tr>
<tr>
<td>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.</td>
<td>8 (36)</td>
</tr>
<tr>
<td>Participants were concerned that the tool would exacerbate existing problems, such as health disparities and alert fatigue.</td>
<td>14 (64)</td>
</tr>
<tr>
<td>Participants were concerned that the tool’s accuracy and implementation were not supported by evidence.</td>
<td>5 (23)</td>
</tr>
<tr>
<td><strong>What changes would increase adoption?</strong></td>
<td></td>
</tr>
<tr>
<td>Adoption of the tool is dependent on its validation, transparency, actionability, and design and could be increased with changes to the interface and usability</td>
<td></td>
</tr>
<tr>
<td>The tool needs to be validated against patient-oriented outcomes so that clinics can quantify the potential return on their investment.</td>
<td>4 (18)</td>
</tr>
<tr>
<td>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.</td>
<td>11 (50)</td>
</tr>
<tr>
<td>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.</td>
<td>6 (27)</td>
</tr>
<tr>
<td>Using user-centered design principles has the potential to minimize the tool’s impact on workflows and maximize readability.</td>
<td>13 (59)</td>
</tr>
<tr>
<td>The ability to customize the tool is important because implementation could differ across practices and clinicians.</td>
<td>2 (9)</td>
</tr>
<tr>
<td>Participants recommended integrating functionality and relevant information from within the EHR(^a).</td>
<td>19 (86)</td>
</tr>
<tr>
<td>Participants recommended other events that could be predicted, including cardiovascular disease, uncontrolled hypertension, worsening depression, care gaps (eg, preventive services), and missed appointments.</td>
<td>22 (100)</td>
</tr>
<tr>
<td><strong>What factors would affect implementation?</strong></td>
<td></td>
</tr>
<tr>
<td>Implementation would require buy-in and need to be tailored to the demands and resources of clinics and communities</td>
<td></td>
</tr>
<tr>
<td>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.</td>
<td>12 (55)</td>
</tr>
<tr>
<td>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.</td>
<td>21 (95)</td>
</tr>
<tr>
<td>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.</td>
<td>20 (91)</td>
</tr>
<tr>
<td>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).</td>
<td>17 (77)</td>
</tr>
<tr>
<td>Successful implementation would require trialability, training, interoperability, and buy-in.</td>
<td>8 (36)</td>
</tr>
</tbody>
</table>

\(^a\)EHR: 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 overwhelmed 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 [incorrect 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 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


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]


Abbreviations

AI: artificial intelligence
EHR: electronic health record
FQHC: federally qualified health center
HbA1c: hemoglobin A1c

©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.