Contents

Original Papers

Sepsis Prediction at Emergency Department Triage Using Natural Language Processing: Retrospective Cohort Study (e49784)
Felix Brann, Nicholas Sterling, Stephanie Frisch, Justin Schrager ................................................................. 3

Learning From International Comparators of National Medical Imaging Initiatives for AI Development: Multiphase Qualitative Study (e51168)
Kassandra Karpathakis, Emma Pencheon, Dominic Cushnan ................................................................. 14

Assessment of ChatGPT-3.5's Knowledge in Oncology: Comparative Study with ASCO-SEP Benchmarks (e50442)
Roupen Odabashian, Donald Bastin, Georden Jones, Maria Manzoor, Sina Tangestanziapour, Malke Assad, Sunita Lakhani, Maritsa Odabashian, Sharon McGee ................................................................. 25

Role of Ethics in Developing AI-Based Applications in Medicine: Insights From Expert Interviews and Discussion of Implications (e51204)
Lukas Weidener, Michael Fischer ................................................................. 33

Identifying Symptoms Prior to Pancreatic Ductal Adenocarcinoma Diagnosis in Real-World Care Settings: Natural Language Processing Approach (e51240)
Fagen Xie, Jenny Chang, Tiffany Luong, Bechien Wu, Eva Lustigova, Eva Shrader, Wansu Chen ................................................................. 48

Beyond the Hype—The Actual Role and Risks of AI in Today's Medical Practice: Comparative-Approach Study (e49082)
Steffan Hansen, Carl Brandt, Jens Søndergaard ................................................................. 59

Jiahui Lu, Huibin Zhang, Yi Xiao, Yingyu Wang ................................................................. 67

Identifying Frailty in Older Adults Receiving Home Care Assessment Using Machine Learning: Longitudinal Observational Study on the Role of Classifier, Feature Selection, and Sample Size (e44185)
Cheng Pan, Hao Luo, Gary Cheung, Huiquan Zhou, Reynold Cheng, Sarah Cullum, Chuan Wu ................................................................. 85

Health Care Professionals' and Parents' Perspectives on the Use of AI for Pain Monitoring in the Neonatal Intensive Care Unit: Multisite Qualitative Study (e51535)
Nicole Racine, Cheryl Chow, Lojain Hamwi, Oana Bucsea, Carol Cheng, Hang Du, Lorenzo Fabrizi, Sara Jasim, Lesley Johannsson, Laura Jones, Maria Laudiano-Dray, Judith Meek, Neelum Mistry, Vibhuti Shah, Ian Stedman, Xiaogang Wang, Rebecca Riddell ................................................................. 97
<table>
<thead>
<tr>
<th>Title</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Reidentification of Participants in Shared Clinical Data Sets: Experimental Study</td>
<td>109</td>
</tr>
<tr>
<td>Privacy-Preserving Federated Survival Support Vector Machines for Cross-Institutional Time-To-Event Analysis: Algorithm Development and Validation</td>
<td>128</td>
</tr>
<tr>
<td>Predictive Performance of Machine Learning–Based Models for Poststroke Clinical Outcomes in Comparison With Conventional Prognostic Scores: Multicenter, Hospital-Based Observational Study</td>
<td>140</td>
</tr>
<tr>
<td>Risk Perception, Acceptance, and Trust of Using AI in Gastroenterology Practice in the Asia-Pacific Region: Web-Based Survey Study</td>
<td>160</td>
</tr>
<tr>
<td>The Impact of Expectation Management and Model Transparency on Radiologists' Trust and Utilization of AI Recommendations for Lung Nodule Assessment on Computed Tomography: Simulated Use Study</td>
<td>173</td>
</tr>
<tr>
<td>Leveraging Machine Learning to Develop Digital Engagement Phenotypes of Users in a Digital Diabetes Prevention Program: Evaluation Study</td>
<td>194</td>
</tr>
<tr>
<td>Machine Learning Methods Using Artificial Intelligence Deployed on Electronic Health Record Data for Identification and Referral of At-Risk Patients From Primary Care Physicians to Eye Care Specialists: Retrospective, Case-Controlled Study</td>
<td>207</td>
</tr>
</tbody>
</table>

**Research Letter**

What Is the Performance of ChatGPT in Determining the Gender of Individuals Based on Their First and Last Names? | 190  |

**Corrigenda and Addenda**

Correction: Using Conversational AI to Facilitate Mental Health Assessments and Improve Clinical Efficiency Within Psychotherapy Services: Real-World Observational Study | 223  |
Sepsis Prediction at Emergency Department Triage Using Natural Language Processing: Retrospective Cohort Study

Felix Brann¹, BSc; Nicholas William Sterling¹, MD, PhD, MS; Stephanie O Frisch¹, PhD, RN; Justin D Schrager¹,², MD, MPH

¹Vital Software, Inc, Claymont, DE, United States
²Department of Emergency Medicine, Emory University School of Medicine, Atlanta, GA, United States

Corresponding Author:
Justin D Schrager, MD, MPH
Department of Emergency Medicine
Emory University School of Medicine
531 Asbury Circle
Annex Building N340
Atlanta, GA, 30322
United States
Phone: 1 404 778 5975
Email: justin@vitaler.com

Abstract

Background: Despite its high lethality, sepsis can be difficult to detect on initial presentation to the emergency department (ED). Machine learning–based tools may provide avenues for earlier detection and lifesaving intervention.

Objective: The study aimed to predict sepsis at the time of ED triage using natural language processing of nursing triage notes and available clinical data.

Methods: We constructed a retrospective cohort of all 1,234,434 consecutive ED encounters in 2015-2021 from 4 separate clinically heterogeneous academically affiliated EDs. After exclusion criteria were applied, the final cohort included 1,059,386 adult ED encounters. The primary outcome criteria for sepsis were presumed severe infection and acute organ dysfunction. After vectorization and dimensional reduction of triage notes and clinical data available at triage, a decision tree–based ensemble (time-of-triage) model was trained to predict sepsis using the training subset (n=950,921). A separate (comprehensive) model was trained using these data and laboratory data, as it became available at 1-hour intervals, after triage. Model performances were evaluated using the test (n=108,465) subset.

Results: Sepsis occurred in 35,318 encounters (incidence 3.45%). For sepsis prediction at the time of patient triage, using the primary definition, the area under the receiver operating characteristic curve (AUC) and macro $F_1$-score for sepsis were 0.94 and 0.61, respectively. Sensitivity, specificity, and false positive rate were 0.87, 0.85, and 0.15, respectively. The time-of-triage model accurately predicted sepsis in 76% (1635/2150) of sepsis cases where sepsis screening was not initiated at triage and 97.5% (1630/1671) of cases where sepsis screening was initiated at triage. Positive and negative predictive values were 0.18 and 0.99, respectively. For sepsis prediction using laboratory data available each hour after ED arrival, the AUC peaked to 0.97 at 12 hours. Similar results were obtained when stratifying by hospital and when Centers for Disease Control and Prevention hospital toolkit for adult sepsis surveillance criteria were used to define sepsis. Among septic cases, sepsis was predicted in 36.1% (1375/3814), 49.9% (1902/3814), and 68.3% (2604/3814) of encounters, respectively, at 3, 2, and 1 hours prior to the first intravenous antibiotic order or where antibiotics were not ordered within the first 12 hours.

Conclusions: Sepsis can accurately be predicted at ED presentation using nursing triage notes and clinical information available at the time of triage. This indicates that machine learning can facilitate timely and reliable alerting for intervention. Free-text data can improve the performance of predictive modeling at the time of triage and throughout the ED course.

(JMIR AI 2024;3:e49784) doi:10.2196/49784

KEYWORDS
natural language processing; machine learning; sepsis; emergency department; triage
Introduction

Background

Sepsis is a life-threatening condition caused by severe infection and dysregulated host response leading to acute organ dysfunction [1]. Affecting 32 million people and contributing to over 5 million deaths per year globally [2], sepsis is a leading cause of death in hospitalizations in the United States and worldwide [3,4]. Early antibiotics have been shown to improve survival [5], while each hour of delayed antibiotic administration has been associated with progressively increased mortality (7.6% increase per hour in septic shock) [6]. Patients who survive sepsis often have long-lasting health and social sequelae [7], and sepsis is ranked among the top 3 most costly conditions to treat in the hospital setting [8]. Accordingly, substantial efforts have been made to identify sepsis early in the hospital course [9]. To date, however, widely used clinical decision support tools that use rule-based methods for detecting sepsis have been limited by low sensitivity and specificity [10,11]. Such tools have been unable to earn clinician trust due to limited accuracy, false positives, and delayed alerts [12]. False positive alerts increase the cognitive load of providers and could expose patients to unnecessary antimicrobials. Moreover, current widely used electronic health record–based sepsis prediction tools have limited performance and often require several hours to elapse to achieve reasonable predictive use [12]. For example, a recent inpatient and intensive care unit (ICU)–based investigation of a commonly used sepsis alerting system showed that although existing systems can generate reasonably accurate sepsis alerts, the median time to notification was 7 hours and, even at that point, accuracy was limited [13]. Taken together, existing clinical decision support systems aimed at detecting sepsis do not provide sufficient accuracy or timeliness of sepsis prediction, resulting in lower adoption due to a lack of clinician trust.

Machine Learning in Sepsis Prediction

Artificial intelligence (AI)–based tools may hold promise to increase the accuracy and timeliness of sepsis prediction, which may allow for earlier delivery of critical interventions such as lifesaving antibiotics. Many of the most promising sepsis predictive algorithms have been limited to use in ICU settings [14], where patients have rich laboratory and imaging data sets and frequent physiologic monitoring. In contrast, accurate prediction of sepsis at initial emergency department (ED) presentation has remained elusive. Until recently, there was a paucity of technology that could make use of the full set of available data, particularly free-text triage notes, at the time of initial ED presentation. A recent study showed that sepsis prediction at the time of triage can be significantly improved using natural language processing (NLP) of free-text data [15].

ED Triage Assessment

When a patient presents to the ED, an initial triage assessment is usually performed by a triage nurse. The triage assessment includes a brief interview of the patient or those accompanying the patient to obtain a reason for presenting to the hospital ED. The content of this interview typically includes a very brief recounting of the patient’s past medical history, relevant medications, family history, and social risk factors. The triage nurse will typically also obtain vital signs (blood pressure, heart rate, temperature, respiratory rate, and oxygen saturation) and pain score. Finally, the triage nurse will assign a patient a triage acuity score. This process usually takes less than 10 minutes. The summation of this encounter is documented in real time, directly after the triage assessment, into the electronic medical record and includes a listing of the vital signs, triage acuity score, and a free-text nursing triage note.

The triage note is recorded into the electronic medical record, typically comprising 1-3 sentences regarding why the patient has presented to the ED and the nurse’s summative impression of this initial assessment. This note is used as a starting point for downstream assessments by providers in the ED. The information contained in the triage note is useful, as it often contains rich data that are difficult to quantify in tabular form. This information is widely used and valued by the clinical staff. However, in its unstructured format, it is not typically used in clinical decision support algorithms and is often unused for several hours until the full provider assessment. We hypothesized that nursing triage notes, combined with other data available at initial ED presentation, could be used to accurately predict sepsis at the time of triage.

Goals of This Investigation

It was previously demonstrated that NLP of nursing triage notes at ED presentation could be used to predict hospital admission and ED resource use [16-18]. In this study, we aimed to demonstrate that an NLP-based model could be used to predict sepsis in adult patients based on the (1) health system sepsis committee and (2) Centers for Disease Control and Prevention (CDC) hospital toolkit for adult sepsis surveillance criteria [1].

Methods

Ethical Considerations

The research study protocol and procedures were reviewed and approved by the institutional review board (STUDY00000099).

Study Design and Setting

A retrospective cohort was constructed using electronic health record data from all 1,234,434 consecutive ED encounters (487,296 unique patients) in 2015-2020 from 4 separate clinically heterogeneous academically affiliated EDs. Hospital A is a community hospital in an urban setting having a patient volume of approximately 65,000 ED visits per year. Hospital B is a community hospital in a suburban setting having a volume of approximately 26,000 visits per year. Hospital C is a quaternary care academic medical setting in a major metropolitan area having an ED patient volume of approximately 48,000 visits per year. Hospital D is a community hospital in a suburban setting having a volume of approximately 36,000 visits per year.

Selection of Participants

Prior studies have suggested that overwhelming viral septicemia during the COVID-19 pandemic led to markedly increased false positive rates of sepsis screening tools [15]. These cases accounted for a substantial portion of ED visits during the initial months of 2020 [19] and led to a sharp decline in ED patient
volume [20]. Accordingly, we excluded encounters (n=94,739) from February 1, 2020, to August 1, 2020, and patients who had a diagnostic code of COVID-19 or positive COVID-19 laboratory test. Patients of 18 years and younger of age were excluded from the study (n=27,238), as defining sepsis in these patients is controversial, and they are often lost to follow-up after they are transferred for admission to pediatric hospitals. Patients whose date of birth or age was not available were also excluded (n=23,434) to ensure that the remaining cohort comprised only adult patients. We subsequently excluded encounters with missing triage notes (n=29,637). The final cohort of interest included 1,059,386 unique clinical encounters.

Sepsis Definition
The primary outcome of sepsis was defined as presumed severe infection and acute organ dysfunction, based on criteria described by the health system sepsis committee. To evaluate model performance against verified sepsis cases, the health system sepsis committee provided physician-reviewed sepsis labels for 7663 patients between June 1, 2019, and October 1, 2019. These cases were oversampled into the test data set. This definition of sepsis was projected onto the remaining data using clinical outcome variables. For sensitivity analyses of model performance, a secondary definition of sepsis was used, based on the US Centers for Medicare & Medicaid Services toolkit criteria [1]. Encounters were counted as sepsis, if they met criteria at any time during the ED course or hospital stay.

Natural Language Processing
NLP techniques have been developed to extract meaning from unstructured free-text data. One such technique is document vectorization. Documents can be transformed into numerical vectors that represent the key information they contain, allowing them to be used by numerical machine learning (ML) techniques.

To generate document embeddings for the nursing triage notes, a distilled BERT (Bidirectional Encoder Representation From Transformers) model pretrained using an unsupervised masked language modeling objective was used as a base. Unlike models pretrained using a causal language modeling objective such as Generative Pre-Train Transformer, which only consider preceding tokens, BERT considers tokens to the right and left of the masked word [21].

The use of large models such as BERT is constrained by the computational resources required for training and inference. DistilBERT [22] is a lighter and faster language model that offers fewer constraints on computational resources, having a depth of only 6 layers, rather than 12, and with token-type embeddings and pooler removed. DistilBERT is trained to replicate the behavior of BERT using “teacher-student” learning, where BERT is the “teacher” and DistilBERT is the “student.” This allows for knowledge distillation in the pretraining phase while retaining 97% of language understanding and being 60% faster.

The base DistilBERT model was fine-tuned using the free textual data from nursing triage notes with the objective of predicting sepsis. We evaluated several pretrained document vectorization models, selecting the optimal one by calculating fine-tuning performance on the training set. Nursing triage notes concatenated with Boolean clinical variables available at the time of triage (ie, high or low vital signs) were then passed through the fine-tuned DistilBERT model to produce document vectors representing the key information they contain. For the document vectors, we selected thresholds for the numeric values based on clinical knowledge and appended text based on the numeric values and those thresholds. Additionally, we developed manual mappings for known clinical abbreviations and converted them into the text. For example, “n/v/d” became “nausea, vomiting, and diarrhea.” The document vectors were then passed through a principal component analysis step to dimensionally reduce them from a length of 768 to 20 components.

Model Training and Testing
For the time-of-triage model, the triage note vectors were combined with other clinical data, such as age, sex, and maximum and minimum vital signs. For the prediction of sepsis after laboratory data availability, a separate comprehensive model was constructed that included the aforementioned variables and additional laboratory data.

While many sepsis indicators have clear unidirectional associations with sepsis risk (ie, heart rate, hypotension, and lactic acid), others can be bidirectional (ie, high or low temperature or white blood cell [WBC] count). In addition, triage note vectors may potentially have complex relationships with sepsis. Accordingly, a decision tree–based technique was chosen for model training over more traditional techniques, such as logistic regression. The combined vectors from the training data set were used to train a decision tree–based ensemble learning model (XGBoost [Extreme Gradient Boosting]) [23] to predict the likelihood of sepsis. The XGBoost model was trained to predict sepsis using the training subset (n=950,921). Model performance was evaluated using the test (n=108,465) subset.

Optimal hyperparameters for the time-of-triage model were determined via grid search. The time-of-triage model was trained using a maximum tree depth of 6, minimum child weight of 15, minimum split loss of 15, learning rate of 0.05, subsample ratio of 0.6, L1 regularization of 0, and L2 regularization of 1. After Bayesian hyperparameter optimization, the comprehensive model was trained using a maximum tree depth of 6, minimum child weight of 13, minimum split loss of 18, learning rate of 0.015, subsample ratio of 0.63, L1 regularization of 0.27, and L2 regularization of 1.87. We accounted for class imbalance by scaling the positive weight parameter to the inverse of the class distribution. Epoch-level evaluation was used to measure model performance during training and identify failing training runs. Heat maps to indicate word and subword importance were generated using the integrated gradients method on the constructed model inputs [24]. Word importance here was calculated on words and subwords returned by the tokenization method.

For analysis of sensitivity, specificity, and false positive rate of the time-of-triage model, a target threshold of model prediction score was selected based on optimizing for a maximal false positive rate of 0.15. For the comprehensive model, we derived...
a classification threshold empirically, based on probability scores, and subsequently applied the threshold to target a maximum false positive rate of 0.1 at 12 hours after ED arrival. The thresholds were selected using model output scores from the training set and were applied to the test data set to evaluate clinical predictive performance metrics. The comprehensive model included known laboratory indicators of sepsis and end organ dysfunction, such as maximum and minimum WBC count, maximum lactic acid, minimum platelets, and maximum bilirubin and creatinine. Comprehensive model performance was evaluated using the test data set at every hour after ED arrival. Model performance was also evaluated at each hospital.

**Sepsis Prediction Prior to the First Intravenous Antibiotic Order**

To estimate how an AI sepsis prediction tool might impact the ordering of antibiotics, we computed the percentage of sepsis encounters that triggered a positive prediction of sepsis prior to antibiotics being ordered or not having antibiotics ordered within the first 12 hours of the encounter. To perform this analysis, we used encounters from the test data set. A dual-model approach was used to emulate sepsis alerting at the time of triage and then subsequently during the ED encounter. Sepsis prediction time was defined as the earlier of either the time-of-triage model or comprehensive model generating a positive prediction of sepsis.

**Evaluation of Model Performances Among Clinically Undetected Sepsis Cases**

To determine how the time-of-triage and comprehensive models may prevent missed sepsis, encounters with sepsis in the test data set were stratified by model prediction of sepsis- versus chart-based indicators of clinical sepsis suspicion. Predictive performance of the model was evaluated among patients who were septic and were or were not screened for sepsis at triage and defined as having either of the following order in less than 30 minutes after time of triage: (1) nursing-driven sepsis screening order set or (2) blood culture.

**Results**

**Characteristics of the Study Patients**

The total data set after exclusions consisted of 1,059,386 unique encounters from 487,296 patients. Sepsis occurred in 35,318 encounters (incidence 3.45%). Median time from arrival to first WBC count collection was 44.9 (IQR 26.2-79.3), 42.8 (IQR 25.6-73.3), and 44.8 (IQR 26.2-79.0) minutes across nonsepsis, sepsis, and all encounters, respectively. Demographic characteristics of the patients are available in Table 1. Gender, race, and temperature were missing in 5.6% (57,082/1,059,386), 13.2% (87,284/1,059,386), and 0.2% (2,034/1,059,386) of encounters, respectively. Respiratory rate, heart rate, oxygen saturation, and blood pressure were missing in 0.1% of encounters. Selected examples of triage notes of encounters where patients were septic are included in Table S1 in Multimedia Appendix 1.
<table>
<thead>
<tr>
<th>Demographic and clinical characteristics of patients across encounters.</th>
<th>Total</th>
<th>Hospital A</th>
<th>Hospital B</th>
<th>Hospital C</th>
<th>Hospital D</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Sepsis(^a), n (%)</strong></td>
<td>1,059,386 (100)</td>
<td>386,961 (36.5)</td>
<td>158,757 (15)</td>
<td>284,794 (26.9)</td>
<td>228,874 (21.6)</td>
</tr>
<tr>
<td>Primary</td>
<td>35,318 (3.3)</td>
<td>9533 (2.5)</td>
<td>3978 (2.5)</td>
<td>12,775 (4.5)</td>
<td>9032 (3.9)</td>
</tr>
<tr>
<td>Secondary</td>
<td>31,542 (3)</td>
<td>9128 (2.4)</td>
<td>3541 (2.2)</td>
<td>12,688 (4.5)</td>
<td>6185 (2.7)</td>
</tr>
<tr>
<td><strong>Age (years), mean (SD)</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>18-24</td>
<td>80,384 (7.6)</td>
<td>35,421 (9.2)</td>
<td>11,466 (7.2)</td>
<td>23,309 (8.2)</td>
<td>10,188 (4.5)</td>
</tr>
<tr>
<td>25-44</td>
<td>344,034 (32.5)</td>
<td>147,085 (38.0)</td>
<td>47,283 (29.8)</td>
<td>91,106 (32.0)</td>
<td>58,560 (25.6)</td>
</tr>
<tr>
<td>45-64</td>
<td>327,584 (30.9)</td>
<td>123,225 (31.8)</td>
<td>53,226 (33.5)</td>
<td>87,113 (30.6)</td>
<td>64,020 (28.0)</td>
</tr>
<tr>
<td>65-74</td>
<td>141,943 (13.4)</td>
<td>44,840 (11.6)</td>
<td>19,709 (12.4)</td>
<td>41,425 (14.5)</td>
<td>35,969 (15.7)</td>
</tr>
<tr>
<td>≥75</td>
<td>165,441 (15.6)</td>
<td>36,390 (9.4)</td>
<td>27,073 (17.1)</td>
<td>41,844 (14.7)</td>
<td>60,137 (26.3)</td>
</tr>
<tr>
<td><strong>Sex, n (%)</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Female</td>
<td>579,798 (57.8)</td>
<td>208,230 (56.8)</td>
<td>90,599 (60.4)</td>
<td>160,710 (59.6)</td>
<td>120,259 (55.6)</td>
</tr>
<tr>
<td>Male</td>
<td>422,506 (42.2)</td>
<td>158,321 (43.1)</td>
<td>59,447 (39.6)</td>
<td>108,611 (40.3)</td>
<td>96,149 (44.4)</td>
</tr>
<tr>
<td><strong>Race, n (%)</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Black</td>
<td>552,432 (50.6)</td>
<td>301,619 (75.6)</td>
<td>35,366 (21.7)</td>
<td>150,454 (51.3)</td>
<td>64,993 (27.6)</td>
</tr>
<tr>
<td>White</td>
<td>380,084 (34.8)</td>
<td>53,427 (13.3)</td>
<td>92,713 (56.8)</td>
<td>104,290 (35.6)</td>
<td>129,654 (56.6)</td>
</tr>
<tr>
<td>Other</td>
<td>39,586 (3.6)</td>
<td>5205 (1.3)</td>
<td>15,827 (9.7)</td>
<td>10,125 (3.5)</td>
<td>8429 (3.6)</td>
</tr>
<tr>
<td>Unreported</td>
<td>87,284 (8.2)</td>
<td>26,710 (6.9)</td>
<td>14,851 (9.4)</td>
<td>19,925 (7.0)</td>
<td>25,798 (11.3)</td>
</tr>
<tr>
<td><strong>Vital signs</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Temperature (°C), mean (SD)</td>
<td>36.8 (0.5)</td>
<td>36.8 (0.5)</td>
<td>36.8 (0.5)</td>
<td>36.7 (0.6)</td>
<td>36.8 (0.5)</td>
</tr>
<tr>
<td>Heart rate (beats per minute), mean (SD)</td>
<td>85.6 (18.8)</td>
<td>86.2 (18.1)</td>
<td>84.5 (18.7)</td>
<td>85.9 (19.1)</td>
<td>84.8 (19.7)</td>
</tr>
<tr>
<td>Systolic BP(^b) (mm Hg), mean (SD)</td>
<td>138.6 (26.7)</td>
<td>138.6 (26.9)</td>
<td>137.6 (24.4)</td>
<td>139.7 (28.9)</td>
<td>137.9 (24.9)</td>
</tr>
<tr>
<td>Diastolic BP (mm Hg), mean (SD)</td>
<td>80.0 (15.5)</td>
<td>80.8 (14.9)</td>
<td>80.2 (14.8)</td>
<td>80.5 (16.0)</td>
<td>77.8 (16.1)</td>
</tr>
<tr>
<td>(\text{SpO}_2) (%) median (IQR)</td>
<td>98.0 (97-100)</td>
<td>98.0 (97-100)</td>
<td>98.0 (97-100)</td>
<td>98.0 (97-100)</td>
<td>99.0 (97-100)</td>
</tr>
<tr>
<td>Respiratory rate (breaths per minute), mean (SD)</td>
<td>18.0 (6.3)</td>
<td>18.2 (6.4)</td>
<td>18.0 (5.9)</td>
<td>18.1 (6.7)</td>
<td>17.9 (5.9)</td>
</tr>
<tr>
<td>Time to first WBC(^d) count (minutes), median (IQR)</td>
<td>44.8 (26.5-80.3)</td>
<td>51.2 (27.3-85.0)</td>
<td>40.9 (20.8-62.8)</td>
<td>47.4 (32.4-90.3)</td>
<td>34.6 (23.1-73.0)</td>
</tr>
</tbody>
</table>

\(^a\)Sepsis primary and secondary definitions based on the health system sepsis committee and Centers for Disease Control and Prevention hospital toolkit for adult sepsis surveillance criteria, respectively.

\(^b\)BP: blood pressure.

\(^c\)\(\text{SpO}_2\): oxygen saturation.

\(^d\)WBC: white blood cell.

### Time-of-Triage and Comprehensive Model Performances

Using the test data set, the time-of-triage model using information available at initial triage for sepsis prediction (primary criteria) demonstrated an area under the receiver operating characteristic curve (AUC) and macro \(F_1\)-score of 0.94 and 0.61, respectively (Figure 1). Sensitivity, specificity, and false positive rate were 0.87, 0.85, and 0.15, respectively. Positive and negative predictive values were 0.18 and 0.99, respectively. Sample model output is available in Figure 2, depicted as heat maps applied to words and subwords of ED nursing triage notes to indicate positive, neutral, or negative contributions to sepsis prediction.
Incorporating data available after initial ED workup, the comprehensive model predicted sepsis based on primary criteria with an initial AUC, sensitivity, and specificity of 0.94, 0.72, and 0.94 at 1 hour after ED arrival, respectively; increasing to an AUC, sensitivity, and specificity of 0.96, 0.87, and 0.91 after 5 hours, respectively; and increasing to AUC, sensitivity, and specificity of 0.97, 0.91, and 0.90 at 12 hours after arrival, respectively (Figure 3). Sensitivity, specificity, and false positive rate at 12 hours were 0.92, 0.89, and 0.11, respectively. Positive and negative predictive values at 12 hours were 0.25 and 0.99, respectively. Similar sepsis prediction results were obtained using the CDC hospital toolkit for adult sepsis surveillance criteria (Table 2) and when stratifying by hospital (Table S2 in Multimedia Appendix 1).
Figure 3. Sepsis predictive performance of the comprehensive model using a test data set, expressed as AUC, at each hour after emergency department arrival. AUC: area under the receiver operating characteristic curve.

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

<table>
<thead>
<tr>
<th>Primary sepsis criteria</th>
<th>Time-of-triage model</th>
<th>Comprehensive model</th>
</tr>
</thead>
<tbody>
<tr>
<td>AUC&lt;sup&gt;a&lt;/sup&gt;</td>
<td>0.94</td>
<td>0.97</td>
</tr>
<tr>
<td>Macro F&lt;sub&gt;1&lt;/sub&gt;</td>
<td>0.61</td>
<td>0.67</td>
</tr>
<tr>
<td>Sensitivity</td>
<td>0.87</td>
<td>0.91</td>
</tr>
<tr>
<td>Specificity</td>
<td>0.85</td>
<td>0.90</td>
</tr>
<tr>
<td>False positive rate</td>
<td>0.15</td>
<td>0.10</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>CDC&lt;sup&gt;b&lt;/sup&gt; hospital toolkit for adult sepsis surveillance</th>
<th>Time-of-triage model</th>
<th>Comprehensive model</th>
</tr>
</thead>
<tbody>
<tr>
<td>AUC</td>
<td>0.92</td>
<td>0.96</td>
</tr>
<tr>
<td>Macro F&lt;sub&gt;1&lt;/sub&gt;</td>
<td>0.57</td>
<td>0.64</td>
</tr>
<tr>
<td>Sensitivity</td>
<td>0.86</td>
<td>0.91</td>
</tr>
<tr>
<td>Specificity</td>
<td>0.83</td>
<td>0.89</td>
</tr>
<tr>
<td>False positive rate</td>
<td>0.17</td>
<td>0.11</td>
</tr>
</tbody>
</table>

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

<sup>b</sup>CDC: Centers for Disease Control and Prevention.

Model Performances Among Clinically Undetected Sepsis Cases

Sepsis screening initiated at triage was defined as having chart-based indicators of sepsis screening ordered within 30 minutes of triage (see Methods section). Within the test data set, there were 3821 encounters having sepsis. Among these, 1671 (43.7%) encounters had sepsis screening initiated at triage. The time-of-triage model accurately predicted sepsis in 76% (1635/2150) of sepsis cases where sepsis screening was not initiated at triage and 97.5% (1630/1671) of cases where sepsis screening was initiated at triage.

Model Performances Among Critical Sepsis Cases

Among patients in the test data set who had sepsis and were ultimately placed on vasopressors or were admitted to the ICU, the time-of-triage model predicted sepsis in 97.9% (329/336) and 91.6% (832/908) encounters, respectively. The comprehensive model predicted sepsis in 100% (336/336) and 95.7% (869/908) encounters, respectively.
Sepsis Prediction Prior to the First Intravenous Antibiotic Order

We retrospectively evaluated the time of sepsis prediction in relation to the first intravenous antibiotic order using a dual-model approach ("time-of-triage" followed by "comprehensive" models). Among septic cases, sepsis was predicted in 36.1% (1375/3814), 49.9% (1902/3814), and 68.3% (2604/3814) of encounters at 3 hours, 2 hours, and 1 hour, respectively, prior to the first intravenous antibiotic order or where antibiotics were not ordered within the first 12 hours.

Model Performance Using Only the First Encounter per Patient

To ensure that model performance was not confounded by past encounters, we performed a sensitivity analysis using only the first encounter per patient in the test data set (n=88,309), excluding subsequent encounters. The time-of-triage model predicted sepsis with an AUC, sensitivity, specificity, and false positive rate of 0.94, 0.85, 0.86, and 0.14, respectively. The comprehensive model predicted sepsis at 12 hours with an AUC, sensitivity, specificity, and false positive rate of 0.97, 0.92, 0.90, and 0.10, respectively.

Analysis of Model Feature Importance

The importance of model features was analyzed by ranking the XGBoost feature importance scores from highest to lowest (Figure S1 in Multimedia Appendix 1). For both the time-of-triage (Figure S2 in Multimedia Appendix 1) and comprehensive (Figure S3 in Multimedia Appendix 1) models, the top features included elements of vital signs (ie, heart rate, temperature, blood pressure, and oxygen saturation) and triage note vectors. For the comprehensive model, the most important features additionally included laboratory metrics such as WBC count, creatinine, and lactic acid.

Discussion

Principal Findings

In this study, data from over 1 million patient encounters across 4 large metropolitan EDs were used to train an NLP-based ML model to detect sepsis at the time of patient presentation to the ED. We demonstrated that free-text nursing triage notes, combined with clinical variables at the time of triage, could be used to accurately predict the occurrence of sepsis at initial ED nursing triage. Moreover, we demonstrated that sepsis could be detected in 76% (1635/2150) of sepsis cases where sepsis screening was not initiated at triage. Finally, the results suggest that AI-based sepsis prediction in the ED may be able to significantly improve the time to antibiotics, which may offer opportunity for lifesaving intervention for patients. Notably, in addition to triage note vectors, the variables with the highest predictive importance were combinations of clinically relevant vital signs (time-of-triage model) and laboratory values, such as WBC count, creatinine, and lactic acid level (comprehensive model). These model characteristics, as well as the ability to map triage note word and subword relative contributions, indicate that the models may offer meaningfully explainable predictions to end users.

To our knowledge, this study is the largest to date to use NLP for sepsis prediction in the ED. We also demonstrated substantially improved accuracy compared to ML-based techniques in prior studies. The ability to incorporate triage notes into an ML model is advantageous for several reasons. First, natural language allows for a broad range of history and examination findings to be compressed into a short free-text note rather than innumerable variables in tabular form. Second, it allows experienced nurses to communicate an overall clinician impression that cannot always be captured by strictly quantitative inputs. In this study, free text from nursing triage notes was used to train a transformer model and was combined as input with other clinical data available at the time of initial triage, with the aim of predicting sepsis. Our findings demonstrate that NLP-based ML models can generate accurate predictions of sepsis at the time of triage and throughout an ED stay. Accordingly, the incorporation of free-text data into models that include data from clinical workups can produce a highly accurate prediction of sepsis.

Importance of Accurate Sepsis Prediction Tools

Existing sepsis alerting systems experience a number of performance difficulties. One of the most widely implemented sepsis detection systems across health systems has been shown to have limited performance due to low sensitivity and precision (33% and 2.4%, respectively). Low predictive performance hinders the clinical use of such systems, despite their aim being to prompt the initiation of lifesaving care. Further impacting their use are high rates of false positive alerts [12]. Increased rates of false positive alerts lead to lower trust among clinicians, alert fatigue and dismissal, and lower adoption [25]. Recently, the incorporation of natural language such as free-text notes into model inputs has been shown to be promising for accurately detecting sepsis as early as during the ED triage process [15].

Prior Studies

To our knowledge, this study is the largest to predict sepsis at the time of ED triage evaluation using NLP-based ML. Ivanov et al [15] reported high predictive performance for sepsis at ED triage with a smaller sample size in 2022. While both this study and Ivanov et al [15] present high sensitivity and specificity and remarkably increased performance compared to traditional screening tools for sepsis, there are important differences between the studies. Whereas Ivanov et al [15] included pediatric encounters, they were excluded in this study, since significantly ill patients of 18 years or younger of age are typically transferred to pediatric hospitals for admission and final diagnoses are unavailable. Accordingly, we excluded these encounters to avoid underestimation of sepsis in the pediatric population, which could have led to type I error with increased reliance on patient age as a predictive feature. A transformer model was also used for the NLP step, which can account for context and surrounding words.

Finally, our approach provides a method to present clinicians with understandable model decision explanations, including heat maps to indicate word importance and contribution to sepsis prediction. We present some examples of these heat maps here. It is important to note that the transformer architecture used in this study assigns meaning using full sentence context, capturing
combined subword and interword relationships, from negation to more complex interactions. As such, these heat maps can be instructive but offer a heavily simplified view of how the algorithm uses triage notes. Additionally, the triage note vectorization is only a part of our complete sepsis algorithm, which also considers additional clinical data throughout the ED encounter.

Limitations

There were several limitations in this study. First, physician-reviewed sepsis labels were only available for a subset of the data and had to be projected onto unlabeled encounters for training purposes using clinical signals. However, model performance was similar when evaluated on the secondary sepsis definition provided in the CDC hospital toolkit for adult sepsis surveillance. Second, the quality of the nursing triage notes is dependent on the clinical skill of the triage nurses, which could vary between EDs. Third, since the COVID-19 pandemic resulted in significant clinical and operational changes, it will be important to include such encounters in future prospective studies. Fourth, no pediatric patients were included, which would bias the model results toward an adult population. Fifth, in this study, it was not possible to detect whether patients were immunocompromised. This is an important subgroup of patients to assess in future studies of ML-based sepsis prediction. Sixth, it was not possible in this study to stratify by causative organism of sepsis, which could affect performance characteristics. Finally, as this study was an investigation of NLP using triage notes, we excluded encounters having missing triage notes.

Conclusions

Using free-text and clinical data available at the time of initial ED triage from over 1 million patient encounters and across 4 hospital-based EDs, we demonstrated that NLP-based ML models are able to achieve high accuracy in predicting sepsis. The implication of these results is that AI-based clinical tools may substantially augment clinician abilities when clinical workup data are sparse, such as at the time of initial ED triage. Since sepsis mortality increases drastically with every passing hour and early clinical intervention is imperative to provide lifesaving treatment, AI-based tools using natural language data, such as free text available in nursing triage notes, may offer critical information to initiate treatment and prevent morbidity and mortality.

Conflicts of Interest

FB, NWS, SOF, and JDS are vice president of data science, machine learning research scientist, director of nursing, and cofounder and chief medical officer, respectively, at Vital Software, Inc, a company engaged in developing artificial intelligence clinical decision support products for the emergency department.

Multimedia Appendix 1
Examples of triage notes, subanalyses, and model explainability.

References


**WBC:** white blood cell

**XGBoost:** Extreme Gradient Boosting

---

©Felix Brann, Nicholas William Sterling, Stephanie O Frisch, Justin D Schrager. Originally published in JMIR AI (https://ai.jmir.org), 25.01.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR AI, is properly cited. The complete bibliographic information, a link to the original publication on https://www.ai.jmir.org/, as well as this copyright and license information must be included.
Learning From International Comparators of National Medical Imaging Initiatives for AI Development: Multiphase Qualitative Study

Kassandra Karpathakis¹, BA, BSc, MPH; Emma Pencheon², BA, MBBS, MSc; Dominic Cushnan³, MBA

¹Decimal.health, Boston, MA, United States
²Foreign, Commonwealth and Development Office, UK Government, London, United Kingdom
³NHS England, London, United Kingdom

Corresponding Author:
Kassandra Karpathakis, BA, BSc, MPH
Decimal.health
50 Milk Street
Boston, MA, 02109
United States
Phone: 1 6086285988
Email: kass.karpathakis@gmail.com

Abstract

Background: The COVID-19 pandemic drove investment and research into medical imaging platforms to provide data to create artificial intelligence (AI) algorithms for the management of patients with COVID-19. Building on the success of England’s National COVID-19 Chest Imaging Database, the national digital policy body (NHSX) sought to create a generalized national medical imaging platform for the development, validation, and deployment of algorithms.

Objective: This study aims to understand international use cases of medical imaging platforms for the development and implementation of algorithms to inform the creation of England’s national imaging platform.

Methods: The National Health Service (NHS) AI Lab Policy and Strategy Team adopted a multiphased approach: (1) identification and prioritization of national AI imaging platforms; (2) Political, Economic, Social, Technological, Legal, and Environmental (PESTLE) factor analysis deep dive into national AI imaging platforms; (3) semistructured interviews with key stakeholders; (4) workshop on emerging themes and insights with the internal NHSX team; and (5) formulation of policy recommendations.

Results: International use cases of national AI imaging platforms (n=7) were prioritized for PESTLE factor analysis. Stakeholders (n=13) from the international use cases were interviewed. Themes (n=8) from the semistructured interviews, including interview quotes, were analyzed with workshop participants (n=5). The outputs of the deep dives, interviews, and workshop were synthesized thematically into 8 categories with 17 subcategories. On the basis of the insights from the international use cases, policy recommendations (n=12) were developed to support the NHS AI Lab in the design and development of the English national medical imaging platform.

Conclusions: The creation of AI algorithms supporting technology and infrastructure such as platforms often occurs in isolation within countries, let alone between countries. This novel policy research project sought to bridge the gap by learning from the challenges, successes, and experience of England’s international counterparts. Policy recommendations based on international learnings focused on the demonstrable benefits of the platform to secure sustainable funding, validation of algorithms and infrastructure to support in situ deployment, and creating wraparound tools for nontechnical participants such as clinicians to engage with algorithm creation. As health care organizations increasingly adopt technological solutions, policy makers have a responsibility to ensure that initiatives are informed by learnings from both national and international initiatives as well as disseminating the outcomes of their work.

(JMIR AI 2024;3:e51168) doi:10.2196/51168

KEYWORDS
digital health; mobile health; mHealth; medical imaging; artificial intelligence; health policy
Introduction

Background
Medical imaging has been identified by many governments as an especially promising application for artificial intelligence (AI) in clinical practice with the potential to enhance disease screening, improve care outcomes, and reduce costs [1-5]. Optimizing AI capabilities requires aggregating and streamlining access to medical imaging data for machine learning (ML) model training and validation and contextualized mechanisms for deployment in clinical workflows.

During England’s National Health Service (NHS) response to the COVID-19 pandemic, the digital health agency (NHSX) created the National COVID-19 Chest Imaging Database (NCCID). The NCCID is a “centralized UK database containing chest X-rays (CXR), Computer Tomography (CT) and Magnetic Resonance Images (MRI) from hospital patients” with COVID-19 [6,7]. It was established to develop, validate, and deploy AI and ML models for supporting the management of patients with severe COVID-19. The creation of the NCCID highlighted the merits and challenges of a centralized approach for collating national imaging data [7].

The NCCID led to a proposal for a generalized national imaging platform for the development, validation, and deployment of AI and ML models in medical imaging. This platform was envisaged to have three technical functions:

1. A data pipeline to facilitate the collection of data nationally
2. A trusted research environment (TRE) to provide access to national data to build and validate new AI and ML products
3. A deployment platform to act as an “app store” for the most up-to-date AI and ML models for users in health care facilities

To support the safe, ethical, and effective creation and deployment of a national imaging platform, the NHS AI Lab developed complementary policy and regulatory initiatives, including a cross-regulatory service to guide developers through the regulation of their AI products [8], understanding of public attitudes toward sharing health data for AI development, and an Algorithmic Impact Assessment tool to identify potential societal impacts of AI products [9].

Beyond understanding the policy and infrastructural requirements, it is important to assess the strengths and weaknesses of such a national approach to produce AI and ML models for imaging that can be deployed in clinical workflows. To make such an assessment, the NHS AI Lab analyzed international efforts to build similar medical imaging platforms in both private and public organizations, some of which were associated with national efforts to diagnose and manage patients with COVID-19. The NHS AI Lab used the outputs of the research to understand the approaches taken and lessons learned and inform the design of England’s national imaging platform.

Objectives
We sought to identify and understand international use cases of and proposals for medical imaging platforms to streamline the innovation-to-deployment journey for health AI models in imaging. We aimed to understand how imaging for AI efforts were structured, identify the constituent parts of the initiatives (eg, technical aspects, users and marketplace, and commercialization), and understand the implications of government policy and regulation. We used this analysis of international use cases to formulate policy recommendations for England’s nascent national AI imaging platform.

Methods

Overview
This research was conducted by NHSX, the former digital health agency and technology policy arm of NHS England. NHSX was merged into the NHS England transformation directorate in 2022. The Strategy and Policy Team at the NHS AI Lab, which was embedded inside NHSX, led and completed the study. This project was conducted between September 2020 and March 2021.

Phase 1: Identification and Prioritization of National AI Imaging Platforms
We conducted a preliminary scan to identify efforts to create national AI imaging platforms in other countries that the NHS AI Lab could analyze in depth.

As the United Kingdom was poised to lead the G7 in 2021, we started with fellow G7 countries: Canada, France, Germany, Italy, Japan, and the United States of America. We then scanned non-G7 countries known within digital health policy circles for their digital health approaches or that had previously responded to an NHSX survey on the use of AI by Global Digital Health Partnership (GDHP) member countries [10]: Australia, Brazil, China, Estonia, Hong Kong, India, Republic of Korea, Rwanda, Singapore, Sweden, Uganda, and Uruguay. Finally, we scanned initiatives in multilateral collaborations (World Health Organization, International Telecommunication Union, and the GDHP) and major private organizations (eg, GE Healthcare and Google).

National AI imaging initiatives were identified by 2 researchers (Abhishek Mishra and EP) through (1) a targeted Google search for each country using [country] and the keywords AI medical imaging platform, medical imaging data, medical AI platform, AI radiology, or COVID-19 medical image AI; (2) a targeted Google search for multilateral collaborations and major private organizations using [name of organization] and the keywords AI medical imaging platform, medical imaging data, medical AI platform, AI radiology, or COVID-19 medical image AI; and (3) a general search on Google, Google Scholar, Twitter, and One HealthTech using the keywords medical imaging AI platform, medical imaging platform, national medical imaging AI platform, or medical imaging AI marketplace. For each search, the first 5 pages of the results were scanned owing to time and resource limitations.

We scored each initiative in comparison with the United Kingdom’s context to prioritize some for the deeper dive in phase 2. Each of the following criteria (n=4) was scored from similar (score=3) to not similar (score=1); initiatives with the
highest total score were deemed most similar to that of the United Kingdom:

1. Similarity of the medical imaging platform to the United Kingdom’s proposed initiative: medical imaging data only versus additional health data, TRE built on top of data to allow for model development, data consolidated in a centralized location or alternative approaches such as federated learning, and parallel building of deployment platform.

2. Size of market: using the country population as a proxy — ≥50 million, 10 to 50 million, and 0 to 10 million.

3. Future trade importance to the United Kingdom: priority markets identified by the NHS Director of AI based on track record of digital health initiatives (note that, at the time of the study, the United Kingdom was the Chair of the G7, and there was strong political interest in the potential for health AI to bolster the United Kingdom’s trade agenda).

4. Regulatory and ecosystem similarity to that of the United Kingdom based on the following: provincial versus national digital health organization, single-payer versus multipayer system, and regulatory approach to AI.

Phase 2: Deep Dive Into National AI Imaging Platforms

For the prioritized initiatives, we conducted a deep dive using the Political, Economic, Social, Technological, Legal, and Environmental (PESTLE) factors framework. PESTLE is a common tool used in policy analysis to gain an overview of an industry [11].

The aims of the deep dive were to (1) identify reliable and robust information to inform the understanding of the international use case; (2) identify hypotheses, gaps, and insights on the AI imaging initiatives for validation during stakeholder interviews; and (3) inform the creation of a deductive framework for the analysis of semistructured interviews. We also identified stakeholders leading AI initiatives to approach for the semistructured interviews in phase 3.

Phase 3: Semistructured Interviews

Semistructured interviews were conducted to understand each prioritized initiative (eg, data used and intended users); its social and political context (eg, regulatory landscape, stakeholders, and public trust), data handling (eg, data and privacy laws), funding sources, and commercialization; and the lessons learned during its development. The discussion guide (Multimedia Appendix 1) was tailored to each country’s unique imaging platform, including the validation of any gaps or insights identified in phase 2.

The interviews were conducted by one principal researcher (KK) with one supporting researcher (EP). Informed consent was obtained from interview participants, and they approved the selected quotes for publication. The interviews lasted up to 1 hour and were audio recorded, and detailed notes were taken. Transcription and translation services were provided by an independent agency. Only one country (Singapore) required the use of translation services to conduct the interview. All other interviews were conducted in English. Both the detailed notes and transcripts from the interviews were analyzed.

The interviews were analyzed using a deductive framework with codes identified from the desk research deep dives (Multimedia Appendix 2). In total, 2 researchers (KK and EP) analyzed each interview independently and compared their coding. Intercoder reliability (ICR) was calculated to assess the reliability of the coding protocol and thematic analysis. ICR was calculated by comparing the level of agreement and disagreement across the coding for 5 pages per transcript [12].

Phase 4: Workshop With NHS AI Lab National Imaging Platform Team

A workshop was conducted with the NHS AI Lab national AI imaging platform team members who were conducting the discovery phase [13]. The workshop aims were to (1) establish top areas of interest from the perspective of the discovery team, (2) explore why these areas are important to the team, and (3) stimulate the discovery teams’ interest in applying the lessons learned from other countries.

The workshop was facilitated by one principal researcher (KK) with one supporting researcher (EP). The workshop lasted 90 minutes, and audio recordings and detailed notes were taken. Participants (n=5) used the web-based Padlet and Jamboard (Google) post-it and “like” functionalities. If required, the researchers noted the participants’ points on their behalf. The workshop audio was transcribed and analyzed.

An overview of the initiatives (n=6) from phase 2 and phase 3 was provided to the attendees using Jamboard. The countries were treated as individual case studies rather than grouped together because of the large degree of heterogeneity between the countries.

A total of 8 themes from the deductive framework were used to guide the workshop: purpose; users; organizational; commercialization; data; incentives; building trust; and law, policy, and regulation. Quotes from the semistructured interviews with stakeholders (phase 3) from each initiative were mapped to the 8 themes for discussion at the workshop.

The nominal group technique was used to identify priority quotes and insights [14]. Participants were asked to vote on the quotes that resonated or were of interest to them using Padlet’s “like” functionality. Each participant had 6 votes per initiative. Voting indicated the discovery team’s priorities and fueled discussions.

The outputs of the deep dives, interviews, and workshop were synthesized thematically into 8 categories with 17 subcategories. The analysis was inspired by a user-centered design insight format [15], which states the context and background, explains the learning, explains the root cause (the why), and explains the motivation behind why the learning has occurred and the ramifications for the NHS AI Lab’s proposed national medical AI imaging initiative.

Phase 5: Formulating Recommendations

The researchers (KK and EP) jointly synthesized all the data gathered from phase 3 to phase 4 to formulate recommendations for the NHS AI Lab national AI imaging initiative. This involved drawing out themes based on the original thematic framework.
identifying learnings pertinent to the United Kingdom, and framing the resulting insights into actionable recommendations.

Final recommendations were presented to the Head of AI Imaging and Director of AI at the NHS AI Lab for consideration. The Head of AI Imaging and the national AI imaging discovery team selected the recommendations that were relevant and actionable for the discovery and future phases of the project. The research team was not privy to this selection.

Ethical Considerations

Internal and external stakeholders were consulted during this policy research and development. Informed consent was obtained from interview and workshop participants. Per NHS’s standard practice, independent ethical review was not required for this research informing policy as it poses negligible risk.

Results

Phase 1: Identified National AI Medical Imaging Platforms

Numerous initiatives (n=34) were identified from preliminary scanning. Most initiatives were country based (21/34, 62%), and the remainder were from major private organizations (10/34, 29%) or multinational organizations (3/34, 9%). Some of the initiatives (7/34, 21%) were prioritized for a deep dive: (1) Digital Health and Discovery Platform (DHDP; Canada), (2) national medical image database (China), (3) Hospital Authority Data Collaboration Laboratory (HADCL; Hong Kong), (4) Research Center for Medical Big Data (Japan), (5) AI Medical Imaging Platform (Singapore), (6) Analytic Imaging Diagnostics Arena (AIDA; Sweden), and (7) Medical Imaging and Data Resource Center (MIDRC; United States).

Phases 2 and 3: Overview of Prioritized National AI Imaging Platforms

In the following sections, we provide a brief overview of each initiative. Multimedia Appendix 3 [16-44] provides a detailed overview of each country’s initiative complemented with findings from the PESTLE analysis and semistructured interviews.

Canada: DHDP

This pan-Canadian initiative was set up to create a nationwide framework to digitally enable research that advances next-generation precision medicine technologies with an emphasis on cancer and improving health outcomes for patients. The DHDP comprises >90 consortium partners spanning academia and the private sector. The initiative focused on numerous types of medical data rather than solely on medical imaging [45] and undertook novel research in federated learning technologies that reflected Canada’s stringent attitudes toward data privacy and sharing.

China: National Medical Image Database

In September 2020, plans were announced for the creation of a standardized national medical image database. The Chinese national medical image database was approved by the National Health Commission [19] to enable hospitals to share patient information and medical images and support the training and development of AI technology for health care. At the time of the study, it was unclear what technology stack the Chinese national imaging database would use and how the initiative would overcome issues of data digitization, cybersecurity, and commercialization.

Hong Kong: HADCL

The HADCL was established to support the formulation of health care policies, facilitate biotechnological research, and improve clinical and health care services. The HADCL is the flexible and interactive data-sharing channel of Hong Kong’s Hospital Authority, with a growing focus on the development of AI and ML algorithms. It is a full-service offering encouraging researchers to partake in collaborative health data projects in a controlled environment using the Hospital Authority’s extensive, longitudinal data [46,47].

Japan: Research Center for Medical Big Data

Japan’s Research Center for Medical Big Data is a platform for AI technology research and development, including a cloud-based platform for hosting medical imaging big data and analyzing medical images. As of 2019, the platform contained >10 million medical images, with participation from at least 60 hospitals. In line with policy at the time of the study, the platform’s primary user base was academia, and projects were for research purposes only.

Singapore: AI-Enabled Medical Imaging Platform

In October 2020, the Integrated Health Information System health laboratory issued a call for collaboration between partners to cocreate an “AI-enabled Medical Imaging Platform” aimed at operationalizing and exploring AI models and applications for medical imaging. The platform will be open and vendor neutral, thereby enabling the deployment of AI models and products from different sources to assist with clinicians’ work.

Sweden: AIDA

AIDA is a dedicated initiative for research and innovation in AI and medical image analysis in Sweden. The initiative brings together academia, health care, and industry to translate innovation into AI-based decision support solutions for imaging diagnostics. The previous mandated creation of national registries containing >5 TB of health data provided the foundation for the AIDA initiative.

United States: MIDRC

The MIDRC is a multi-institutional initiative established in response to the COVID-19 pandemic. The aim was to foster ML innovation through the sharing of imaging and associated clinical data regarding COVID-19 [48]. At the time of the study, agreements for sharing relevant medical imaging data were in the process of being signed with several sites, but no data were being hosted on the platform.

Phases 3 and 4: Derived Themes and Insights

Stakeholders (n=16) representing 7 initiatives were approached for interviews. Stakeholders (n=13) from 6 initiatives accepted the interview invitations (13/16, 81% acceptance rate). The stakeholders from participating countries were 38% (5/13) from Canada, 23% (3/13) from Japan, 8% (1/13) from Hong Kong, 2% (1/13) from the United States, 23% (3/13) from Sweden, and 8% (1/13) from Japan.
8% (1/13) from Singapore, 8% (1/13) from Sweden, and 15% (2/13) from the United States. Stakeholders from China (3/16, 19%) did not respond to the request for an interview.

For the interview coding, the ICR between the researchers (KK and EP) was calculated to be 0.41, indicating moderate reliability [12,49]. The outputs of the deep dives, interviews, and workshop were synthesized thematically into categories (n=8) with subcategories (n=17).

Multimedia Appendix 4 presents the categories, subcategories, and corresponding thematic synthesis within each of the other countries’ initiative including key insights, quotes, and learnings.

Phase 5: Recommendations

Overview

We provided 12 recommendations for the NHS AI Lab’s proposed national AI imaging platform. Each recommendation is grounded in the themes and insights from phase 2 to phase 4 (see Multimedia Appendix 4). The corresponding themes for each recommendation are also provided.

Narrative

Recommendation 1: The NHS AI Lab develop a purposeful narrative of why and how a national medical imaging initiative is necessary, outlining what health needs it will meet and supporting this with demonstration of its benefit and potential

Developing a strong value proposition should be married with demonstrable benefit. The narrative should be cross-cutting, speaking not only to purpose but also to trust and incentives, with transparency regarding the drivers of the initiative. Previous work by the NHS AI Lab on behalf of the GDHP has also argued that countries should take a “needs based” approach to AI-driven technology development to create both maximal benefit on health outcomes and foster buy-in and support from stakeholders and the public [47,50].

A purposeful narrative for the NHS AI Lab’s national medical imaging initiative will support interdisciplinary collaboration and ensure long-term political, financial, and social support for the initiative based on a clear understanding of its importance and utility to the health system. An important aspect of this narrative is to reference the value of the initiative as a social or public good that creates public value [51].

The corresponding themes for this recommendation are (A) demonstrable benefit of the initiative, (B) health system needs as the primary driver, (C) community and shared purpose, and (O) transparency and communication. transparency and communication.

Recommendation 2: The NHS AI Lab moves away from the language of “platform” to talking about the national medical imaging initiative as an “initiative” and community space for growing the United Kingdom’s understanding and ability to use AI in medical imaging

The United Kingdom’s national medical imaging “initiative” should be carefully framed, using language that reflects what is offered and conveys mindset and purpose. The connotations of “national” in the initiative name given the involvement (or lack thereof) of the Devolved Administrations (DAs) should be considered. In addition, the NHS AI Lab should develop an approach for involving the DAs.

The corresponding themes for this recommendation are (C) community and shared purpose and (D) embracing and enabling the central role of health care professionals.

Users and Service Offering

Recommendation 3: The NHS AI Lab develops wraparound services to maximize engagement and capitalize on the expertise of varied users; by removing the need to technically upskill in AI development while also providing opportunities for users to do so if they wish, the initiative can broaden participation and avoid disincentivizing users with different and valuable areas of specialty

The NHS AI Lab should invest in wraparound services, specifically offering tools and professional technical skills that are tailored to fill a gap that users, such as health care professionals, have when it comes to developing AI. It appears from international comparators that the main draw and success has not been the platform itself but the supportive services to enable users to engage, collaborate, and develop AI-driven technologies regardless of their technical expertise. Examples include but are not limited to clinical fellowships on health data, networking or pairing clinicians with data scientists, training courses on what is AI and how to develop models, and low-code and no-code AI model development tools. The NHS AI Lab should explore opportunities to build these wraparound services from existing programs in the digital health ecosystem.

The corresponding themes for this recommendation are (D) embracing and enabling the central role of health care professionals, (E) recognizing that users are not discrete groups, and (F) importance of wraparound services.

Recommendation 4: The NHS AI Lab continues to embrace interdisciplinary work while designing, developing, and implementing the national medical imaging initiative; the inherent tensions and perspectives between disciplines are needed to deliver on health system needs

Interdisciplinary work is central to harnessing the breadth of expertise required to build and sustain an initiative that truly addresses health system needs. This means embracing the central role of health care professionals and ensuring the participation of people who have a system view of health and social care, as well as those with frontline experience who will be the ultimate end users of any AI products developed on the platform. Prioritizing user-centered design and health care professionals’ experience means that technical expertise must take an important facilitative and instructive role to both guide and learn from health care professionals about how to leverage AI-driven technologies in the health system. By facilitating interdisciplinary work, radiologists’ expertise can be applied to shore up the quality and appropriateness of the imaging data used. We recommend that active steps be taken to foster collaborative working relationships across disciplines drawing
on lessons for interdisciplinary collaboration outlined by Blandford et al [52] and on the examples of activities run in Sweden and Japan.

The corresponding themes for this recommendation are (B) health system needs as the primary driver for AI development, (D) embracing and enabling the central role of health care professionals, and (E) recognizing that users are not discrete groups.

**Sustainability and Future-Proofing**

**Recommendation 5: The NHS AI Lab consider the financial sustainability of the national medical imaging initiative from the outset and how this maps to the proposed commercial model**

All the international comparators who did not have a clear commercial model raised concerns about financial sustainability. It is worth bearing in mind that demonstrable benefit does not guarantee enduring government support with respect to funding. We recommend that the NHS AI Lab national medical imaging initiative considers how the work will be sustained beyond current funding and ensures that options for commercialization are not excluded by virtue of how the initiative is designed (ie, data-sharing arrangements that preclude commercialization). For the NHS AI Lab’s national medical imaging initiative to have longevity, it is important to keep as many commercial options on the table as possible, including generating revenue from certain aspects of the initiative and exploring public-private partnerships. This could include providing data subsets to fulfill specific needs, such as validation, that can be commercialized as a distinct offering.

The corresponding themes for this recommendation are (I) ensure financial sustainability, (J) differing or absent commercial models, and (L) subsetting data offerings.

**Recommendation 6: The NHS AI Lab continues to explore different commercial models for the national medical imaging initiative with a focus on how it might commercialize aspects of the initiative rather than taking an all-or-none approach**

Commercialization is likely necessary to ensure the financial sustainability of the initiative. Commercial models were an afterthought for many international comparators, who conveyed the sense that commercialization was viewed as being in opposition to the public good. We recommend thinking about commercial options early on, not only from a practical perspective of building the initiative with this in mind but also to construct a narrative that can interweave commercialization and private sector involvement with the public good. The NHS AI Lab should continue working with internal teams (ie, the NHSX Centre for Improving Data Collaboration) to ensure that the NHS gains fair value for the public from commercial arrangements.

The corresponding themes for this recommendation are (I) ensure financial sustainability, (J) differing or absent commercial models, and (L) subsetting data offerings.

**Recommendation 7: The NHS AI Lab explore and potentially adopt some of the future-proofing mechanisms used by international comparators**

Sweden and the United States exemplified ways to future-proof data-sharing mechanisms, including specific clauses in data-sharing agreements that granted them the power to revoke data access or extend it to future offerings. This is important for safeguarding against issues further down the road and streamlining the process of setting up data-sharing agreements. Sweden was cognizant that currently, anonymized data might become reidentifiable with advances in data analysis and wanted to mitigate this risk from the outset through the ability to revoke access at any time. We also recommend that, if and where possible, the initiative infrastructure is future-proofed and reusable so that it will be fit for purpose in years to come and offer benefits to other similar initiatives.

The corresponding themes for this recommendation are (M) future-proofing mechanisms for data sharing and (N) a focus on public and social good.

**Recommendation 8: The NHS AI Lab balances the need to deliver at pace with the up-front investment of time and effort required to ensure that the resulting initiative is sustainable and future-proofed**

A variety of pressures to deliver at pace were identified by international colleagues, which at times nudged countries toward “kicking the can down the road” when it came to thorny challenges such as commercialization. Although a certain level of pace is necessary to demonstrate benefit and garner support, this should be tempered to ensure an up-front investment of time and effort that delivers sustainable returns.

The corresponding themes for this recommendation are (A) demonstrable benefit of the national medical imaging initiative and (G) tempering the pace of development.

**Recommendation 9: The NHS AI Lab consider under what conditions it would be acceptable and feasible to move beyond human-in-the-loop approaches in the national medical imaging initiative’s resultant AI-driven technologies**

All countries maintained the need for a human to be “in the loop” to ensure the safety, accountability, and acceptability of AI development and products. Human-in-the-loop refers to models that require human interaction, whereby human oversight can intervene and determine the outcome of a process or event. However, there is an undertone that moving beyond human-in-the-loop approaches is the future state of AI-driven technology in health and care (in some conditions, not yet defined). We recommend that the NHS AI Lab start considering not only the safety and accountability of systems without humans and when this would be deemed appropriate but also the public perception of not having unique or individualized care.

The corresponding themes for this recommendation are (K) common and continuing data challenges, (O) transparency and communication, and (P) keeping humans in the loop.
Recommendation 10: The NHS AI Lab accounts for the environmental impact of the national medical imaging initiative and establishes how it aligns with a sustainable health and social care system

No international comparators had considered the environmental impact of their initiative or how they were positioned in relation to delivering a sustainable health and care system. This presents an opportunity for the United Kingdom to lead in this domain considering the health system needs not only for now but also for the future. We recommend that the NHS AI Lab develop an understanding of how the national medical imaging initiative could affect both positively and negatively an economically and environmentally sustainable health system. This is an important element of future-proofing the work and ensuring that it is fit for purpose in the coming decades (note: the NHS AI Lab strategy team has started considering how AI could contribute to the NHS goal of reaching net zero by 2045 and to an environmentally sustainable health and care system [53]).

The corresponding themes for this recommendation are (B) health system needs as the primary driver and (N) a focus on public and social good.

Policy and Regulation

Recommendation 11: The NHS AI Lab leverage its privileged position as the guiding health technology organization within both the civil service and the NHS to continue advocating and driving policy and regulatory change: the United Kingdom’s national medical imaging initiative is a tangible use case for uncovering the issues and providing examples of how they could be solved

All countries recognized that their current policies and regulations were not fit for the purpose of AI development and implementation in clinical settings. There was a range of mindsets regarding how to balance operating within constraints and advocating to change them. The NHS AI Lab is uniquely positioned within the government to drive the necessary changes in the United Kingdom making use of existing collaborations with regulatory bodies and DAs. We recommend that the national medical imaging initiative, with clearly articulated and demonstrable benefits to the health system, be used as evidence for this advocacy work.

The corresponding themes for this recommendation are (H) building on existing infrastructure and resources and (Q) advocating for policy, regulatory, and legal frameworks that are fit for purpose.

Recommendation 12: The NHS AI Lab leverage the work already undertaken in validation of AI models as a unique selling point for the United Kingdom’s national medical imaging initiative

No international comparators had progressed to the deployment and widespread adoption of AI-driven technologies developed through their initiatives. One of the bottlenecks for this is a clear validation process, an area in which the NHS AI Lab is well placed to take the lead given the existing work that has been done in this domain. We recommend that this is capitalized on as a unique selling point for the national medical imaging initiative to demonstrate an innovation funnel that runs smoothly through to the deployment of assured technologies.

The corresponding themes for this recommendation are (H) building on existing infrastructure and resources and (Q) advocating for policy, regulatory, and legal frameworks that are fit for purpose.

Discussion

Principal Findings

The NHS AI Lab sought to learn from countries developing medical imaging platforms to streamline the innovation-to-deployment journey for AI and ML algorithms for medical imaging. The research team conducted secondary and primary research with use cases from multiple countries to develop a deep understanding of the approaches for structuring a medical imaging platform program, how to set up supportive policy and regulatory initiatives, and form relationships with international stakeholders.

In addition to providing 12 recommendations for the NHS AI Lab to implement, the research team identified five areas in which the NHS AI Lab could offer a unique value proposition:

1. Galvanizing the already operating proof of concept, the NCCID program, to demonstrate benefit and secure stable United Kingdom government funding and support.
2. Within the new medical imaging platform, build in the ability to validate AI and ML algorithms as well as deploy them in health care settings. Only a few international initiatives built in the ability to validate algorithms and create a deployment pipeline, which is crucial for ensuring the effectiveness of algorithms during implementation.
3. Create wraparound offerings tailored to researchers, developers, and private companies operating in the United Kingdom. This may include tools to facilitate the creation of algorithms, training and workshops for upskilling, computational power, legal and regulatory support, and demand signaling for areas of clinical specialty in which there is high demand for AI and ML development.
4. Consider the environmental impact and sustainability of the medical imaging platform and the resultant carbon output from the outset.
5. Publicly demonstrate that the NHS AI Lab has incorporated collaborative international learnings and best practices.

Strengths

The primary strength of the project was the NHS AI Lab’s openness to learning from other countries. Throughout our engagement with selected countries (Canada, Hong Kong, Japan, Singapore, Sweden, and the United States), we established that no other initiative had conducted international landscaping to inform strategy and implementation. Our work highlights the benefit of not reinventing the wheel in health AI initiatives but reaching out to build on the experience and expertise of others.

Second, the internal discovery team responsible for designing and building the NHS AI Lab’s medical imaging platform was engaged throughout the delivery of this project. Their engagement culminated in the workshop to elicit feedback and
prioritize insights, followed by the selection of final recommendations. Often, policy and strategy research is conducted before or separately from the team creating and building a product. Policy and strategy research conducted in isolation may not provide practical and usable recommendations that can be taken forward during product development.

Limitations

We identified 3 key limitations of this project. First, no literature review was conducted to inform the research. Owing to the novelty of creating medical imaging platforms for AI development, we instead decided to conduct a scan of potential international efforts via targeted Google, Google Scholar, and social media searches.

Second, the ICR reliability indicates some variation in coding assignments between the 2 researchers (KK and EP). Coding variability could be attributed to (1) the level of experience analyzing qualitative research and (2) the depth of understanding of the topics discussed by the interview participants. It is important to note that the resultant ICR of 0.41 indicates moderate reliability, which falls within tolerance as outlined by Landis and Koch [49] and O’Connor and Joffe [12].

Third, the study did not delve into the role and importance of postmarket monitoring or surveillance. In some interviews, it appears that this topic was not top of mind as they were working on initiatives that were in the beginning stages and algorithms were not yet actively deployed into the market for clinical use. However, since the completion of this project, the NHS AI Lab has funded the United Kingdom Medicines and Healthcare products Regulatory Agency to deliver several work packages, including updating legislation to require more robust postmarket surveillance for software as a medical device [54].

Conclusions

Policy makers and digital developers internationally are chasing the potential for AI and ML algorithms to transform health care, with medical imaging seen as low-hanging fruit for realizing this ambition. Algorithms in health care are not confined to national borders, so how this ambition is realized by each country is particularly important. This paper outlines work undertaken by the NHS AI Lab to ensure that the investment in and creation of a generalized national medical imaging platform for the innovation and deployment of AI and ML algorithms in England is informed by international experience.

Acknowledgments

First, the authors would like to thank the stakeholders from each initiative for participating in this research. The authors learned a lot from each and every one of them and value their contributions. Second, the authors would like to thank the NHS AI Lab at NHS England, formerly at NHSX, for supporting the publication of this policy research and embedding the recommendations into the decision-making process for England’s national imaging platform efforts. Finally, the authors would like to acknowledge Abhishek Mishra, who supported the earlier stages of the research while in a PhD intern placement at the NHS AI Lab and was funded by a Wellcome Trust doctoral scholarship. All research was conducted by staff members employed by or deployed to NHSX. No external funding was received to conduct the research. DC, Director of AI at the NHS AI Lab, NHS England, is the guarantor of the publication.

Authors’ Contributions

KK conceptualized and supervised all stages of this project, including securing project resources, data curation, and project administration. DC was the main NHSX stakeholder and lead for the conceptualization and development of the National COVID-19 Chest Imaging Database and national artificial intelligence imaging platform. KK developed the research methodology with input from Abhishek Mishra and conducted this research alongside Abhishek Mishra and EP. EP and KK developed the discussion guide and deductive thematic analysis coding framework for the semistructured interviews. KK was the lead interviewer, and EP was the second interviewer and notetaker. KK and EP developed the workshop materials. KK was the lead workshop facilitator with support from EP. Transcription and translation services were provided by Prestige Network. KK and EP completed the thematic analysis and data synthesis. KK wrote the first draft of the manuscript. All the authors contributed to the drafting and editing of the manuscript and have approved the final version.

Conflicts of Interest

KK and EP were working at NHSX at the time of the study. DC was employed at NHSX at the time of the study and at NHS England at the time of writing.

Multimedia Appendix 1

Template discussion guide.

[DOCX File, 19 KB - ai_v3i1e51168_app1.docx]

Multimedia Appendix 2

Deductive thematic and coding framework.

[DOCX File, 34 KB - ai_v3i1e51168_app2.docx]
Multimedia Appendix 3
Description of international initiatives.
[DOCX File, 42 KB - ai_v3i1e51168_app3.docx]

Multimedia Appendix 4
Thematic synthesis.
[DOCX File, 207 KB - ai_v3i1e51168_app4.docx]

References


Abbreviations
- AI: artificial intelligence
- AIDA: Analytic Imaging Diagnostics Arena
- DA: Devolved Administration
- DHDP: Digital Health and Discovery Platform
- GDHP: Global Digital Health Partnership
- HADCL: Hospital Authority Data Collaboration Laboratory
- ICR: intercoder reliability
- MIDRC: Medical Imaging and Data Resource Center
- ML: machine learning
- NCCID: National COVID-19 Chest Imaging Database
- NHS: National Health Service
- PESTLE: Political, Economic, Social, Technological, Legal, and Environmental
- TRE: trusted research environment

©Kassandra Karpathakis, Emma Pencheon, Dominic Cushnan. Originally published in JMIR AI (https://ai.jmir.org), 04.01.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR AI, is properly cited. The complete bibliographic information, a link to the original publication on https://www.ai.jmir.org/, as well as this copyright and license information must be included.
Assessment of ChatGPT-3.5's Knowledge in Oncology: Comparative Study with ASCO-SEP Benchmarks

Roupen Odabashian¹, MD; Donald Bastin², MD; Georden Jones³*, PhD; Maria Manzoor, MD; Sina Tangestanianiour, MD; Malke Assad⁴, MD; Sunita Lakhani⁵, MD; Maritsa Odabashian³,6, Bsc (c); Sharon McGee⁷,8*, MD, PhD

¹Department of Oncology, Barbara Ann Karmanos Cancer Institute, Wayne State University, Detroit, MI, United States
²Department of Medicine, Division of Internal Medicine, The Ottawa Hospital and the University of Ottawa, Ottawa, ON, Canada
³Mary A Rackham Institute, University of Michigan, Ann Arbor, MI, United States
⁴Department of Plastic Surgery, University of Pittsburgh Medical Center, Pittsburgh, PA, United States
⁵Department of Medicine, Division of Internal Medicine, Jefferson Abington Hospital, Philadelphia, PA, United States
⁶The Ottawa Hospital Research Institute, Ottawa, ON, Canada
⁷Department of Medicine, Division of Medical Oncology, The Ottawa Hospital and the University of Ottawa, Ottawa, ON, Canada
⁸Cancer Therapeutics Program, Ottawa Hospital Research Institute, Ottawa, ON, Canada

*these authors contributed equally

Corresponding Author:
Roupen Odabashian, MD
Department of Oncology, Barbara Ann Karmanos Cancer Institute, Wayne State University
4100 John R St
Detroit, MI, 48201
United States
Phone: 1 (313) 745 3000 ext 7731
Fax: 1 313 576 8120
Email: roupen.odabashian@mclaren.org

Abstract

Background: ChatGPT (Open AI) is a state-of-the-art large language model that uses artificial intelligence (AI) to address questions across diverse topics. The American Society of Clinical Oncology Self-Evaluation Program (ASCO-SEP) created a comprehensive educational program to help physicians keep up to date with the many rapid advances in the field. The question bank consists of multiple choice questions addressing the many facets of cancer care, including diagnosis, treatment, and supportive care. As ChatGPT applications rapidly expand, it becomes vital to ascertain if the knowledge of ChatGPT-3.5 matches the established standards that oncologists are recommended to follow.

Objective: This study aims to evaluate whether ChatGPT-3.5’s knowledge aligns with the established benchmarks that oncologists are expected to adhere to. This will furnish us with a deeper understanding of the potential applications of this tool as a support for clinical decision-making.

Methods: We conducted a systematic assessment of the performance of ChatGPT-3.5 on the ASCO-SEP, the leading educational and assessment tool for medical oncologists in training and practice. Over 1000 multiple choice questions covering the spectrum of cancer care were extracted. Questions were categorized by cancer type or discipline, with subcategorization as treatment, diagnosis, or other. Answers were scored as correct if ChatGPT-3.5 selected the answer as defined by ASCO-SEP.

Results: Overall, ChatGPT-3.5 achieved a score of 56.1% (583/1040) for the correct answers provided. The program demonstrated varying levels of accuracy across cancer types or disciplines. The highest accuracy was observed in questions related to developmental therapeutics (8/10; 80% correct), while the lowest accuracy was observed in questions related to gastrointestinal cancer (102/209; 48.8% correct). There was no significant difference in the program’s performance across the predefined subcategories of diagnosis, treatment, and other (P=.16, which is greater than .05).

Conclusions: This study evaluated ChatGPT-3.5’s oncology knowledge using the ASCO-SEP, aiming to address uncertainties regarding AI tools like ChatGPT in clinical decision-making. Our findings suggest that while ChatGPT-3.5 offers a hopeful outlook for AI in oncology, its present performance in ASCO-SEP tests necessitates further refinement to reach the requisite competency levels. Future assessments could explore ChatGPT’s clinical decision support capabilities with real-world clinical
scenarios, its ease of integration into medical workflows, and its potential to foster interdisciplinary collaboration and patient engagement in health care settings.

(JMIR AI 2024;3:e50442) doi:10.2196/50442

KEYWORDS
artificial intelligence; ChatGPT-3.5; language model; medical oncology

Introduction

OpenAI released ChatGPT, a pioneering artificial intelligence (AI) language model, in late 2022. ChatGPT-3 is an AI chatbot that can comprehend user input and react to it in a manner that is natural and human-like [1]. The program was trained on a large body of data sourced from the internet, including textbooks, articles, social media posts, and web-based forums, up to the last quarter of 2021 [2]. It works by analyzing user input text to generate a response using a probabilistic distribution of words and phrases derived from its training data. To date, it has significantly impacted numerous disciplines, including law, health care, and medical education [3-6]. Large language models like ChatGPT-3.5 represent a significant advancement in the preceding class of deep learning–based models, by facilitating the interpretation, processing, and production of natural language [7].

The use of AI has rapidly emerged as a promising approach in the health care industry, where it has been applied to medical imaging analysis, drug discovery, and patient monitoring [8]. Recent research has evaluated ChatGPT-3.5’s abilities to respond to standardized questions from professional examinations for law and the United States Medical Licensing Examination (USMLE) [3,4]. ChatGPT-3.5 was able to achieve passing grades on these examinations while providing logical and informative explanations. Additionally, studies have been conducted to assess ChatGPT’s capabilities in responding to international medical licensing examinations from countries such as Italy, France, Spain, the United Kingdom, and India. The success rates observed ranged between 22% and 73% [9].

AI and ChatGPT showcase substantial promise in augmenting medical consultations, offering preliminary diagnostic suggestions, and providing a vast knowledge base for medical practitioners and patients alike [10]. However, while it embarks on a path toward a more integrated health care AI system, several limitations hinder its full potential. The model’s reliance on historical data without the ability to access real-time patient data can lead to outdated or inaccurate information dissemination. Additionally, its inability to comprehend nuanced human emotions and the ethical implications surrounding patient data privacy remain significant hurdles [11].

AI has displayed a notable deficiency in grasping context and nuance, elements that are fundamental for delivering safe and effective patient care [12]. Furthermore, analyzing the prospects of job automation in health care, Frey and Osborne [13] have projected that while administrative roles within the sector, such as health information technicians, exhibit a high likelihood of automation at 91%, the odds plummet to a mere 0.42% for the automation of roles held by physicians and surgeons. This stark contrast underscores the intricate nature of medical practice, which extends beyond the mere application of codified knowledge. Additionally, there is a burgeoning discussion around the ethical dimensions of using conversational AI in medical practice. The crux of the issue revolves around the substantial volume of high-quality data required to train these models. Present-day algorithms are often honed on biased data sets, inheriting not just the availability, selection, and confirmation biases inherent in the data but also displaying a propensity to exacerbate these biases [14]. Looking ahead, the evolving capabilities of AI hint at the potential for tackling more sophisticated tasks, such as orchestrating experiments or future clinical trials [15] or engaging in peer review processes [16].

The American Society of Clinical Oncology Self-Evaluation (ASCO-SEP) program created a comprehensive educational program to help physicians keep up to date with the many rapid advances in the field. The question bank consists of multiple choice questions (MCQs) addressing the many facets of cancer care, including diagnosis, treatment, and supportive care. It is intended to evaluate participants’ knowledge and give them feedback to direct future learning. The program is largely regarded as the leading resource for cancer specialists seeking to gain and maintain professional licensure in the field of medical oncology [17].

However, the evolving complexities of oncological care necessitate additional tools that can aid oncologists in clinical decision-making. By assessing ChatGPT-3’s ability to answer ASCO-SEP questions, this study’s objective is to understand ChatGPT’s potential to serve as a supportive instrument in clinical decisions, offering instantaneous insights for health care providers, and to identify novel and efficient educational aids in oncology, with a specific emphasis on their role in clinical decisions.

Methods

Input Data

Questions were sourced from ASCO-SEP, which consists of approximately 1000 MCQs covering the spectrum of cancer care. The question bank was accessed from February 2023 to March 2023. As ChatGPT-3.5 can only generate responses to textual data, the study excluded questions with images, tables, or other non-textual content. Questions consisted of an information stem followed by a specific question with 3-5 possible answers (A-E), along with their corresponding letter choices, only 1 of which was correct. Figure 1 illustrates the workflow for data sourcing, input, encoding, and analysis.
Before proceeding with the analysis, a random spot check was performed. For this, a random subset of the ASCO-SEP questions was selected, and their answers, explanations, or related content were manually cross-referenced with Google’s index to ensure that they were not present before January 1, 2022, the last date accessible to the ChatGPT training data.

During this study, we used the free version of ChatGPT-3.5. At that time, ChatGPT-4 and its associated plugins were not yet available.

**Encoding**
We imported individual ASCO-SEP questions, including the information stem and multiple-choice response options, into the ChatGPT-3.5 interface. The questions were formatted to include the question stem, followed by each potential response on a separate line. We did not change the structure of the questions given to ChatGPT-3.5 and entered them in the original format provided by ASCO-SEP without altering the phrasing or the wording. A new conversation session was started in ChatGPT for each question. We did not provide ChatGPT-3.5 with any prompts and offered only one opportunity to answer each question.

**Data Analysis**
Selected questions were grouped by cancer type or discipline (eg, breast, lung, and colon cancer) with further subcategorization based on content such as treatment, diagnosis, or other. ChatGPT was deemed to have responded correctly if it chose the correct answer as defined and provided by ASCO-SEP. The study team did not define or determine the correct answer. The program was not asked to provide justifications or references for answers. No point was assigned if ChatGPT-3.5 provided an answer that was not from the options given. Questions where ChatGPT-3.5 chose 2 possible answers or chose multiple answers and did not commit to a single best answer were also considered wrong, even if 1 of the responses was correct.

For statistical analysis, data were logged, scored, and analyzed in Excel (Microsoft Corp). Specifically, a chi-square test was performed to determine if there was a significant difference in the distribution of correct answers across different categories or groups.

**Results**
A total of 1040 questions were extracted from the ASCO-SEP question bank.

The questions covered 15 cancer types or disciplines. The largest portion focused on breast (223/1040; 21.4%) and gastrointestinal (209/1040; 20%) cancers, with ≤1% (13/1040) covering central nervous system malignancies, developmental therapeutics, and prevention/epidemiology (Table 1).
Table 1. Question distribution and proportions by cancer type or specialty area.

<table>
<thead>
<tr>
<th>Cancer type or discipline</th>
<th>Number of questions (N=1040), n (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Breast cancer</td>
<td>223 (21.4)</td>
</tr>
<tr>
<td>Gastrointestinal cancer</td>
<td>209 (20)</td>
</tr>
<tr>
<td>Thoracic oncology</td>
<td>137 (13.1)</td>
</tr>
<tr>
<td>Hematological malignancies</td>
<td>121 (11.6)</td>
</tr>
<tr>
<td>Genitourinary cancer</td>
<td>97 (9)</td>
</tr>
<tr>
<td>Melanoma and skin cancer</td>
<td>43 (4)</td>
</tr>
<tr>
<td>Sarcoma</td>
<td>36 (3)</td>
</tr>
<tr>
<td>Head and neck</td>
<td>36 (3)</td>
</tr>
<tr>
<td>Gynecologic cancers</td>
<td>36 (3)</td>
</tr>
<tr>
<td>General oncology</td>
<td>29 (3)</td>
</tr>
<tr>
<td>Supportive and palliative care</td>
<td>28 (3)</td>
</tr>
<tr>
<td>Genetics and genomics</td>
<td>17 (2)</td>
</tr>
<tr>
<td>Central nervous system</td>
<td>13 (1)</td>
</tr>
<tr>
<td>Developmental therapeutics</td>
<td>10 (1)</td>
</tr>
<tr>
<td>Prevention and epidemiology</td>
<td>5 (0.5)</td>
</tr>
</tbody>
</table>

Varying levels of accuracy were observed in ChatGPT-3.5’s performance in answering questions based on different cancer types or disciplines (Table 2). The highest accuracy was achieved in questions related to developmental therapeutics (8/10; 80% correct), while the lowest accuracy was observed for questions related to gastrointestinal cancer (102/209; 48.8% correct).

Table 2. Accuracy rates by cancer type or specialty area.

<table>
<thead>
<tr>
<th>Cancer type or discipline</th>
<th>Discipline-specific accuracy rates, n/N (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Developmental therapeutics</td>
<td>8/10 (80)</td>
</tr>
<tr>
<td>Central nervous system</td>
<td>10/13 (77)</td>
</tr>
<tr>
<td>Melanoma and skin cancer</td>
<td>28/43 (65)</td>
</tr>
<tr>
<td>Genetics and genomics</td>
<td>11/17 (65)</td>
</tr>
<tr>
<td>General oncology</td>
<td>18/29 (62)</td>
</tr>
<tr>
<td>Gynecologic cancers</td>
<td>22/36 (61)</td>
</tr>
<tr>
<td>Supportive and palliative care</td>
<td>17/28 (61)</td>
</tr>
<tr>
<td>Prevention and epidemiology</td>
<td>3/5 (60)</td>
</tr>
<tr>
<td>Head and neck</td>
<td>21/36 (58)</td>
</tr>
<tr>
<td>Breast cancer</td>
<td>130/223 (58.3)</td>
</tr>
<tr>
<td>Sarcoma</td>
<td>20/36 (57)</td>
</tr>
<tr>
<td>Thoracic oncology</td>
<td>77/137 (56)</td>
</tr>
<tr>
<td>Hematological malignancies</td>
<td>66/121 (55)</td>
</tr>
<tr>
<td>Genitourinary cancer</td>
<td>49/97 (51)</td>
</tr>
<tr>
<td>Gastrointestinal cancer</td>
<td>102/209 (48.8)</td>
</tr>
<tr>
<td>Total</td>
<td>583/1040 (56.1)</td>
</tr>
</tbody>
</table>

Questions were further subcategorized as “diagnosis,” “treatment,” and “other,” with the latter covering topics such as biostatistics, cancer staging, and treatment complications. Out of the total questions, 73.1% (760/1040) were related to cancer treatment, 10% (99/1040) focused on diagnosis, and the remaining 17.4% (181/1040) were categorized as “other” (Table 3). Accuracy based on subcategory also varied, with 55% (418/760) of treatment questions, 63% (62/99) of diagnosis questions, and 56.9% (103/181) of “other” questions answered correctly (Table 2). There was no significant difference in the program’s performance across the predefined subcategories of
diagnosis, treatment, and other ($P=.16$, which is greater than .05).

<table>
<thead>
<tr>
<th>Category</th>
<th>Number of questions, n (%)</th>
<th>Overall accuracy, n/N (%)</th>
<th>$P$ value$^a$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Treatment</td>
<td>760 (73.1)</td>
<td>418/760 (55)</td>
<td>.16</td>
</tr>
<tr>
<td>Diagnosis</td>
<td>99 (10)</td>
<td>62/99 (63)</td>
<td>.16</td>
</tr>
<tr>
<td>Other</td>
<td>181 (17.4)</td>
<td>103/181 (56.9)</td>
<td>.16</td>
</tr>
<tr>
<td>Overall</td>
<td>1040 (100)</td>
<td>583/1040 (56.3)</td>
<td>.16</td>
</tr>
</tbody>
</table>

$^a$Chi-square test.

Overall, ChatGPT-3.5 achieved a score of 56.3% (583/1040) for correct answers provided across all categories. Of note, responses were marked as incorrect if ChatGPT-3.5 provided 2 or more answers, even if 1 of those answers was correct (37/1040, 3%; Figure 1).

**Discussion**

**Overview**

In this study, we evaluated the performance of ChatGPT-3.5 in answering ASCO-SEP questions designed for medical oncologists in training and practice to support licensure and ongoing medical education. To facilitate a fair and rigorous assessment, spot checks were performed to ensure answers were not present in the program training data, and questions were entered in separate sessions to avoid grounding bias. Furthermore, questions were presented in their original format, as seen by physicians, with no changes made to prompt the program.

Over 1000 questions were posed to the program, spanning the spectrum of cancer care, with an overall score of 56.3% (583/1040) achieved. While promising, this is, however, below the accepted threshold of 70% that is required by ASCO-SEP to claim CME credits using their question bank [18].

Since the launch of ChatGPT-3.5, several studies have evaluated the program’s performance on medical examinations. A notable study conducted by Kung et al [3] assessed ChatGPT-3.5’s performance on the USMLE taken by US medical students. The results showed that ChatGPT-3.5 performed at, or near, the passing threshold for all 3 examinations. Specifically, the accuracy rates for USMLE Steps 1, 2 CK, and 3 were 68.0%, 58.3%, and 62.4%, respectively, which are acceptable passing scores. Gilson et al [19] reported similar results, where ChatGPT-3.5 scored 60% on USMLE test questions. It is worth noting that although the authors used questions published on the USMLE website after the training date cutoff for ChatGPT, which is late 2021, many of these questions were similar to those published in previous years. Moreover, these questions were discussed on web-based forums, which may explain the higher scores achieved [20]. Additionally, previous studies have evaluated ChatGPT-3.5’s performance in microbiology [21] and pathology [22] and have shown promising outcomes in these fields with an accuracy rate of 80%.

Several factors might explain why ChatGPT’s performs differently on USMLE compared to ASCO-SEP questions. First, the ASCO-SEP is tailored for medical oncologists, delving deep into cancer care, while USMLE caters to a broader set of medical students, covering general medical knowledge. Given that ChatGPT-3.5’s training data spans a wide range of topics, it’s plausible that the content aligns more with the generalized medical queries of USMLE than the specialized focus of ASCO-SEP. Additionally, the structure and phrasing of questions play a critical role, potentially influencing AI’s response accuracy. The questions within the USMLE typically features keywords that assist students in selecting an answer from the provided options. Conversely, the ASCO-SEP presents more specialized questions, challenging physicians’ ability to discern first- and second-line treatments for a specified condition [23]. For instance, in 1 of the numerous subreddits [24] available web-based that was likely included in ChatGPT’s training data set [25] students discuss how certain keywords aid them in answering examination questions. These data might have assisted ChatGPT in responding to USMLE questions in a previous paper that tested ChatGPT’s performance on the USMLE [3,19]. However, such keywords are not used or discussed among physicians engaging with ASCO-SEP questions.

There are additional possible explanations for the observed performance of ChatGPT-3.5 in this study. One key factor is the comprehensive data set of over 1000 questions used, which allowed for a more thorough and holistic evaluation of the program’s performance compared to previous studies [3,19,26,27]. Another contributing factor may be the dynamic and rapid scientific and clinical advances that occur in the field of oncology, which ChatGPT-3.5 could not fully tackle given that its training data is limited to pre-2022 internet data, with restricted access to key databases in the field like PubMed [28].

ChatGPT-3.5 demonstrated varying levels of accuracy in answering questions across the different cancer types and disciplines. Questions related to developmental therapeutics had the highest accuracy rate (80%, 8/10); however, the limited question sample size may not have allowed a complete assessment. Indeed, ChatGPT-3.5’s lowest score was achieved in gastrointestinal cancer, which contained one of the largest numbers of questions in the bank (102/209, 48.8%), suggesting that broader assessments may identify more knowledge gaps. This study did not, however, find any significant difference in ChatGPT-3.5’s performance across the subcategories of diagnosis, treatment, and others.

While ChatGPT-3.5 is not yet fully dependable for complex decision-making in medical oncology, it shows promise in the
field. In recent years, we have witnessed significant progress in neural networks, and the future of health care is becoming increasingly multimodal. Oncologists now rely on more than just text-based information when prescribing treatments. They consider a wide range of factors, including diverse image types, genomic data, and social determinants of health. However, in the past, developing multimodal machine learning models seemed like an overly ambitious goal. Thankfully, the landscape has changed, and we have seen exciting advancements in this area through various publications in 2022 and 2023 [29,30]. These studies have showcased the potential applications of multimodal models in the field of oncology, bringing us closer to a more comprehensive and holistic approach to cancer care.

Based on its performance in this study, we do not think that AI can aid oncologists in clinical decision-making at this time. However, it may excel in other tasks in the field [31]. Experts might look to language-generating AI to reduce the burden on humans who create questions and explanations for tests. However, it should be noted that ChatGPT-3.5 is not a useful tool without human supervision at this point, given its potential to fabricate references that may sound plausible but are incorrect [14,32,33]. Oncologists can also use it for administrative tasks such as drafting notes [34] or crafting communication messages for patients [11]. Additionally, while a previous study by Johnson et al [35] demonstrated that ChatGPT can be used by patients to answer common cancer myths and questions, the questions used in this study were already featured on the National Cancer Institute’s webpage and were likely part of ChatGPT’s training data [25] and fewer questions were used. We can infer from this study that the answers provided by ChatGPT still require review by an oncologist to ascertain their accuracy.

In the future, AI has the potential to assist oncologists in critical aspects such as determining optimal chemotherapy dosages [36] and aiding in diagnostics within fields like radiology and pathology [37]. By leveraging the capabilities of these advanced language models, health care professionals can access valuable insights and support in making informed decisions regarding treatment plans. Moreover, patients can also reap the advantages of AI-driven technologies by receiving more accurate diagnoses and tailored treatment approaches, ultimately leading to improved outcomes and enhanced patient care [38].

This study does, however, have several important limitations. First, as ASCO-SEP only consists of MCQs, we did not challenge ChatGPT-3.5 with any other question formats (eg, open-ended), which may have yielded different results. Furthermore, MCQs may not fully reflect the complexity of clinical scenarios that oncologists face in their practice. Second, we did not test the variability of the answers provided by ChatGPT. Each question was presented to ChatGPT 3.5 only once, and the first answer was scored given that previous studies showed high consistency of ChatGPT answers [39]. Finally, we could have performed a qualitative assessment of ChatGPT-3.5 answers to gain insights into the etiology of its errors as a guide to future required improvements.

Conclusions

In conclusion, this study explored the capacity of ChatGPT-3.5’s knowledge in medical oncology using the ASCO-SEP. We aimed to bridge the knowledge gaps surrounding the efficacy of AI-driven tools like ChatGPT-3.5 in supporting clinical decision-making. Our assessment revealed that while ChatGPT-3.5 shows promise for the future of AI in oncology, its current performance on ASCO-SEP underscores a pressing need for further refinement to meet the competency standards in this complex field. Future evaluations of ChatGPT could extend to assessing its capability in clinical decision support, gauging its accuracy in real-life clinical scenarios, and its ease of integration into medical workflows. Evaluating GPT-4 as a resource to aid oncologists in clinical decision-making, an aspect not available during the tenure of this study, could significantly contribute to the field. The tool’s facilitation of interdisciplinary collaboration among health care professionals and its impact on patient engagement and communication are other potential areas of investigation.

Conflicts of Interest

None declared.

References


10. Asch D. An interview with ChatGPT about health care. NEJM Catal Innov Care Deliv 2023;4(2) [FREE Full text]


Abbreviations

AI: artificial intelligence
ASCO-SEP: American Society of Clinical Oncology Self-Evaluation Program
MCQ: multiple choice question
USMLE: United States Medical Licensing Examination
Role of Ethics in Developing AI-Based Applications in Medicine: 
Insights From Expert Interviews and Discussion of Implications

Lukas Weidener¹, BSc, Dr med; Michael Fischer¹, PhD

Research Unit for Quality and Ethics in Health Care, UMIT TIROL – Private University for Health Sciences and Health Technology, Hall in Tirol, Austria

Corresponding Author:
Lukas Weidener, BSc, Dr med
Research Unit for Quality and Ethics in Health Care
UMIT TIROL – Private University for Health Sciences and Health Technology
Eduard-Wallnöfer-Zentrum 1
Hall in Tirol, 6060
Austria
Phone: 43 17670491594
Email: lukas.weidener@edu.umit-tirol.at

Abstract

Background: The integration of artificial intelligence (AI)–based applications in the medical field has increased significantly, offering potential improvements in patient care and diagnostics. However, alongside these advancements, there is growing concern about ethical considerations, such as bias, informed consent, and trust in the development of these technologies.

Objective: This study aims to assess the role of ethics in the development of AI-based applications in medicine. Furthermore, this study focuses on the potential consequences of neglecting ethical considerations in AI development, particularly their impact on patients and physicians.

Methods: Qualitative content analysis was used to analyze the responses from expert interviews. Experts were selected based on their involvement in the research or practical development of AI-based applications in medicine for at least 5 years, leading to the inclusion of 7 experts in the study.

Results: The analysis revealed 3 main categories and 7 subcategories reflecting a wide range of views on the role of ethics in AI development. This variance underscores the subjectivity and complexity of integrating ethics into the development of AI in medicine. Although some experts view ethics as fundamental, others prioritize performance and efficiency, with some perceiving ethics as potential obstacles to technological progress. This dichotomy of perspectives clearly emphasizes the subjectivity and complexity surrounding the role of ethics in AI development, reflecting the inherent multifaceted nature of this issue.

Conclusions: Despite the methodological limitations impacting the generalizability of the results, this study underscores the critical importance of consistent and integrated ethical considerations in AI development for medical applications. It advocates further research into effective strategies for ethical AI development, emphasizing the need for transparent and responsible practices, consideration of diverse data sources, physician training, and the establishment of comprehensive ethical and legal frameworks.

KEYWORDS
artificial intelligence; AI; medicine; ethics; expert interviews; AI development; AI ethics

Introduction

Background

Artificial intelligence (AI) has been considered a key technology in medical advancement for several years [1]. Recent developments in AI, exemplified by the broad availability and widespread use of advanced AI-based chat applications, such as ChatGPT, have underscored the capabilities of technology [2]. This study specifically focuses on AI-based applications in medicine, highlighting the importance of ethics in their development, with an emphasis on the role of developers. Considering the inherent complexities associated with AI and its applications in medicine along with the multifaceted nature of AI ethics, this introduction aims to provide a comprehensive foundation for this publication.
Artificial Intelligence

Early definitions of AI, such as by McCarthy et al [3], primarily focused on the potential for machines to simulate all facets of human intelligence: "...the basis of the conjecture that every aspect of learning or any other feature of intelligence can in principle be so precisely described that a machine can be made to simulate it." Newer definitions, such as the one from the European Parliament, expand this scope and describe AI as “the ability of a machine to display a range of humanlike capabilities, including reasoning, learning, planning, and creativity,” encompassing a broader spectrum of intelligent behaviors [4].

Following the evolving definitions of AI, the term broadly encompasses various technologies, each with unique characteristics and applications. The scientific community commonly categorizes these technologies as “strong AI” and “weak AI” [5]. “Strong AI” refers to systems whose cognitive capabilities are comparable with human intelligence across a wide range of tasks and contexts [5]. However, most current applications, particularly in medicine, are categorized as “weak AI.” This category includes systems designed to perform specific tasks using cognitive abilities comparable with those of humans but within a limited scope [6]. Within the category of “weak AI,” 2 primary subfields are prominent: expert systems and machine learning (ML) [6]. Expert systems, categorized under “symbolic AI,” operate based on predefined rules and instructions set by human experts [7]. In contrast, ML represents the “statistical AI” subfield [8]. ML focuses on pattern recognition within large data sets, enabling the system to learn and make predictions or decisions based on the data [9]. A notable example of such advancements in “statistical AI” is the development of large language models, such as ChatGPT, which demonstrate the evolving capabilities of AI in understanding and generating humanlike text, offering new possibilities, and raising unique ethical considerations in their application [10].

Despite the significant technological advances in the field of AI and, in particular, “weak AI,” “strong AI” which would entail cognitive abilities on par with human intelligence across diverse areas, remains largely theoretical with no substantial application in medicine to date [11]. Therefore, “weak AI” will be the foundation of this publication, specifically focusing on the development and associated ethical considerations of “symbolic AI” and “statistical AI” applications in medicine.

AI in Medicine

The technological advancements and capabilities of AI in medicine, as exemplified by a range of AI-based applications such as ML algorithms and expert systems, are anticipated to transform various aspects of health care, such as diagnostics or personalized treatment planning [1].

For example, ML algorithms, a key subset of “statistical AI,” are of particular interest in medicine because of their capability to analyze large data sets, including a wide array of medical images such as x-rays, magnetic resonance imaging, computed tomography, and dermatological photographs [8]. In radiology, ML algorithms enhance image interpretation by identifying the features associated with specific pathologies. For instance, in mammography, ML assists radiologists in detecting microcalcifications and subtle changes in the breast tissue, which may indicate the early stages of breast cancer [12]. Similarly, in dermatology, ML-powered tools analyze photographic data of skin lesions and moles, thereby providing critical diagnostic insights [13]. By distinguishing between benign and malignant lesions with high accuracy, the early detection of skin cancer can be improved. The integration of ML in image-based diagnostics can not only enhance diagnostic accuracy but also have the potential to speed up the diagnostic process [8]. This reduction in analysis time leads to quicker diagnostic outcomes, enabling earlier intervention and treatment, which are crucial for improving patient care [14].

Expert systems in medicine, a subfield of “symbolic AI,” are primarily exemplified by Clinical Decision Support Systems (CDSS) [15]. By leveraging predefined rules and knowledge from medical experts, these systems can provide recommendations for diagnosis and therapy options, potentially enhancing the decision-making process in clinical settings [16]. CDSS often use information from various sources, such as electronic health records, patient history, and latest medical research, to offer evidence-based suggestions. In addition to offering diagnostic and treatment guidance, CDSS can play a significant role in identifying potential adverse drug events, which is a critical aspect of patient safety [16]. By cross-referencing a patient's current medications with the proposed treatments, CDSS can alert health care providers to possible drug-drug interactions, allergic reactions, or contraindications based on the patient’s medical history or known conditions [15].

In addition to diagnostic and decision support applications, AI contributes to other areas of medicine, such as medical research and drug development. In medical research, AI algorithms are used to analyze complex information, such as genetic, environmental, and lifestyle data, which can be used for personalized medical approaches, enabling more targeted therapies based on individual patient profiles [17]. Furthermore, AI can be used to identify potential therapeutic compounds more quickly and efficiently than traditional methods [18]. AI systems can simulate and predict how different compounds interact with biological targets, thereby reducing the time and cost of drug trials. This capability is particularly crucial in rapidly responding to emerging global health challenges, such as the development of vaccines and treatments for new diseases [18]. Furthermore, although AI-based chat applications, such as ChatGPT, have not been specifically developed for use in medicine, they possess extensive medical knowledge, making their potential application in various medical contexts a subject of increasing interest [2]. Although advancements in the field of AI can offer transformative benefits for medicine, they also introduce new ethical considerations and challenges that warrant attention [19,20].

AI Ethics

AI ethics can be defined as “a set of values, principles, and techniques that employ widely accepted standards of right and wrong to guide moral conduct in the development and use of AI technologies” [21]. Although this definition does not specifically focus on or include the field of medicine, it
emphasizes the importance of values and principles in the development of AI technologies. In medicine, the fundamental principles of medical ethics formulated by Beauchamp and Childress—autonomy, nonmaleficence, beneficence, and justice—are of paramount influence and relevance [22].

The principle of autonomy emphasizes respecting patients’ rights to make informed decisions regarding their own health. In the context of AI-based applications in medicine, the principle of autonomy often refers to the development of technologies that support and enhance patient decision-making while maintaining transparency, explainability, and accountability [23,24]. This also refers to the development of AI-based applications that not only provide accurate diagnostic and treatment recommendations but also present their findings in a manner that is understandable and useful for both patients and health care professionals. The principle of nonmaleficence, emphasizing the commitment to do no harm, has become increasingly important in the context of growing role of AI in health care. Adhering to this principle requires the establishment of stringent safety protocols and comprehensive testing of AI technologies to prevent unintended consequences, such as biases in decision-making that could lead to misdiagnosis or unequal treatment of patients [24].

Bias in AI systems, particularly in medical applications, is a significant concern. For instance, ML algorithms used in image-based diagnostics, such as those used in radiology or dermatology, may develop biases based on the data they are trained on [25]. If these algorithms are primarily trained on data sets that lack diversity, they might be less accurate in diagnosing conditions in patient populations that are underrepresented in the training data [26]. This can lead to disparities in diagnostic accuracy and effectiveness, potentially harming certain groups of patients [16,27]. Similarly, in CDSS, which rely on predefined rules and medical knowledge, there is a risk of inherent biases being transferred into the system. If the input data or rules within these systems reflect historical biases or unequal treatment practices, the CDSS might perpetuate these issues, leading to recommendations that are not equitable or appropriate for all patients [16].

Addressing the challenges related to autonomy and nonmaleficence is fundamental for ensuring that AI in medicine aligns with the principles of beneficence and justice. The principle of beneficence, or acting in the best interests of the patient, emphasizes that AI-based applications in medicine should be developed with the primary goal of improving patient outcomes and enhancing quality of care [23]. Finally, the principle of justice requires that AI technologies in health care promote fairness and equity. This means ensuring equitable access to the benefits of AI advancements regardless of a patient’s socioeconomic status or background [24].

In light of these ethical principles, the role of developers in creating AI-based applications in medicine has become critically important. Developers bear a particular responsibility to ensure that the design and implementation of these technologies adhere to the ethical standards outlined by autonomy, nonmaleficence, beneficence, and justice [28]. A deep understanding and awareness of the ethical implications during the development process are essential, as the principles and guidelines frequently discussed in the current literature should be integrated from the early stages of AI application development [29,30]. This integration is not just theoretical but requires practical implementation and consistent consideration throughout the development process of AI-based applications in health care [31]. Despite the crucial role that developers play in embedding these ethical principles into AI technologies, there remains a gap in the literature regarding how developers perceive and prioritize ethics in their work [32,33]. Addressing this gap is essential for ensuring the responsible development and use of AI in medicine and aligning technological advancements with the core values of medical ethics.

**Objective**

The field of AI-based medical applications is rapidly advancing; however, a significant gap remains in understanding how ethical considerations are integrated into this development process. Recognizing the frequent calls in the literature for consistent inclusion of ethics in AI development, this study aimed to bridge this gap by exploring the perceptions, priorities, and conflicts related to ethics among AI experts. Specifically, this study sought to answer the following questions:

- How do AI experts perceive the role of ethics in the development of AI-based medical applications?
- How do AI experts perceive the relationship between ethical considerations and the technical development of AI-based applications in medicine?

The primary objective of this study is not only to answer these critical questions but also to provide an in-depth discussion of the results, particularly focusing on the associated ethical implications. This exploration is vital for understanding how ethical considerations can be more effectively integrated into the development of AI technologies in medical settings with the aim of contributing to the responsible and beneficial advancement of this field.

**Methods**

To address the study’s objective, a secondary analysis of the exploratory expert interviews was performed using qualitative content analysis. These interviews were initially conducted to explore the essential knowledge and understanding of AI in medicine, with the aim of specifying teaching content on AI for medical education [34].

**Ethical Considerations**

Ethics approval was granted by the Research Committee for Scientific Ethical Questions of the UMIT TIROL—Private University for Health Sciences and Health Technology, Hall in Tirol, Austria, for both the initial data collection and secondary analysis of the data relevant to this study (approval number: 3181; January 16, 2023).

The methodology and reporting of the research findings in this study were guided by the Standards for Reporting Qualitative Research to ensure clarity and transparency [35].
Expert Characteristics

Of the 12 experts included in the primary research study, 7 met the inclusion criteria for this study and provided information relevant to the study objective. For this secondary analysis, individuals were defined as experts if they had been engaged in the research or practical development of AI-based applications in medicine for at least 5 years. In this regard, 4 experts were involved in the development of AI-based applications as part of their research activities (eg, researchers at the German Research Center for Artificial Intelligence, professor for medical informatics), such as enhanced AI-assisted imaging. The remaining 3 experts were primarily engaged in the practical development of various AI-based applications for use in medicine (eg, voice recognition in hospitals or assistance in diagnosis in medical practices) as part of their main professional activities in the private sector (eg, software development). Additional inclusion criteria were sufficient language skills (German) and consent for the transcription of the interviews and their evaluation. All 7 participating experts were situated and working in Germany, providing a national perspective on the development of AI in medicine. Of the 7 experts included in this secondary analysis, 6 identified as male and 1 (E2) identified as female. Although all experts met the inclusion criteria of being engaged in research or the practical development of AI-based applications in medicine for at least 5 years, 3 experts (E1, E2, and E4) had more than 10 years of professional experience in the relevant field. In addition, 3 experts had more than 15 years of experience in the field of research and practical development of AI-based applications (E3, E5, and E7). Table 1 presents a detailed overview of the experts’ characteristics included in this study.

Table 1. Characteristics of the experts included in the secondary analysis.

<table>
<thead>
<tr>
<th>Expert number</th>
<th>Professional position</th>
<th>Domain of expertise</th>
</tr>
</thead>
<tbody>
<tr>
<td>E1</td>
<td>Research and development (AI)</td>
<td>Machine learning in pathology</td>
</tr>
<tr>
<td>E2</td>
<td>Data scientist</td>
<td>AI in radiology</td>
</tr>
<tr>
<td>E3</td>
<td>Senior software developer</td>
<td>Clinical Decision Support Systems</td>
</tr>
<tr>
<td>E4</td>
<td>Research and development (AI)</td>
<td>AI in cancer diagnosis</td>
</tr>
<tr>
<td>E5</td>
<td>Professor for medical informatics</td>
<td>Natural language processing in medicine</td>
</tr>
<tr>
<td>E6</td>
<td>Data scientist</td>
<td>AI-assisted voice analysis for diagnosis</td>
</tr>
<tr>
<td>E7</td>
<td>Senior software developer</td>
<td>Clinical Decision Support Systems</td>
</tr>
</tbody>
</table>

*AI: artificial intelligence.

Data Collection

In the initial data collection phase of the primary study, experts were recruited primarily via email. In addition, participants were asked to recommend other potential experts for the interviews, thereby expanding the recruitment network. This direct recommendation approach enabled the inclusion of 2 additional experts in the primary study. Before the interviews were recorded, the experts were informed about the study and the associated data protection regulations during recruitment and at the beginning of the interviews. All interviews were conducted using a video service provider (Cisco Webex Meetings) and were recorded on an audio basis (manual recording via an analog dictation device; average interview length 34.02, SD 4.1 minutes).

To ensure the protection of all collected and generated data, they were stored offline on a password-protected storage device in a lockable cabinet, with access limited to the researcher. The anonymized data will be stored for 10 years following the date of collection to enable reproducibility and deleted after to ensure confidentiality. All participating experts explicitly consented to both the initial analysis and the use of their data for future research purposes, as in the case of this study.

Data Analysis

The expert interviews were transcribed using the transcription software f4transcript and anonymized according to the transcription rules of Dresing and Pehl [36]. The evaluation of the collected data was conducted with software support (QCAmap, version 1.2.0; Microsoft Excel, version 16.66) and was rule based according to the methodology of qualitative content analysis by Mayring (inductive procedure) [37]. Relevant categories were defined directly from the material and were controlled or revised after viewing 40% of the material. After defining the categories, the entire material was reviewed, and relevant text passages were assigned to the respective main and subcategories.

The interviews were conducted and analyzed in German. For this publication, all identified and relevant text passages were translated into the English language. The primary research team conducted the initial translation, followed by a review and revision by a professional academic translator.

It is noteworthy that the data analysis in this study was guided by the research team’s perspective and understanding of ethics. As such, the interpretation of the data and subsequent conclusions are shaped by the team’s affiliation with the research unit for quality and ethics in health care. Consequently, ethical considerations, particularly in health care and medicine as well as in the development and application of AI technologies in these fields, are considered important. The emphasis on ethics should be considered when interpreting the results of this study.

Furthermore, the aspect of theoretical saturation in this secondary analysis warrants detailed discussion. Given its distinct objectives, this study selectively used interviews with 7 of the 12 experts, chosen based on the specific inclusion
criteria of engagement in research or practical development of AI-based applications in medicine for over 5 years. The remaining 5 experts from the primary study, who primarily focused on teaching and research without a direct emphasis on developing AI-based applications for medicine, did not meet the inclusion criteria for this secondary analysis. This selection, inherent to the secondary nature of the data, led to a focused but relatively limited breadth in certain areas, resulting in incomplete saturation in the 2 subcategories. Specifically, the subcategories of “Data Protection” (section Subcategory 3: Data Protection) and “Demands” (section Subcategory 3: Data Protection) demonstrated incomplete saturation, each substantiated by only a single reference. In contrast, theoretical saturation for the other categories can be assumed, given the multiple references that support the established themes and the lack of new insights, suggesting the need for additional categories.

Acknowledging this limitation is crucial, particularly in the context of future research opportunities aimed at more comprehensively exploring these underrepresented areas. However, the reliability of the results extends beyond the theoretical saturation. It is also underscored by the expertise and extensive experience of the participating experts, each with at least 5 years of AI research or practical development in medicine. Their profound insights, combined with the systematic and iterative analysis methodology, ensured that the extracted themes were representative and comprehensive, despite the gaps noted in certain subcategories. Consequently, although the findings in the “Data Protection” and “Demands” categories might benefit from further exploration in future studies, the current analysis offers a robust and insightful understanding of the primary themes related to ethical considerations in AI development for medical applications.

To ensure detailed and comprehensive data collection, a semistructured interview guideline was used for primary data collection. This interview guideline included questions directly related to the study’s objectives and incorporated both immanent and exmanent questioning. Reflecting the research team’s focus on ethics in health care and medicine, the semistructured interview guidelines incorporated 2 questions directly relevant to the study’s objectives:

- How do you perceive the role of ethics in the context of AI-based medical applications?
- What are your experiences with ethical considerations and the development of AI-based applications in medicine?

In addition to the 2 questions directly addressing the objective of this study, an interview guideline was constructed to promote openness by emphasizing the immanent and exmanent questions. Examples of the questions used are as follows:

- You have mentioned the challenge of integrating ethics into AI development. Could you elaborate on the specific ethical considerations you find most relevant in this context?
- In your view, who should bear responsibility for the ethical issues in AI-based applications—users or developers?

Using both direct and immanent as well as exmanent question types, the interviews aimed to provide an in-depth exploration of the topic of AI in medicine, including the development of AI-based applications for use in medicine.

**Results**

**Overview**

On the basis of the qualitative content analysis of the expert interviews, 3 main categories with 7 subcategories were defined using anchor examples. **Textbox 1** provides an overview of the main categories and subcategories defined.

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

<table>
<thead>
<tr>
<th>Essential foundation</th>
</tr>
</thead>
<tbody>
<tr>
<td>• Awareness</td>
</tr>
<tr>
<td>• Consequences</td>
</tr>
<tr>
<td>• Data protection</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Results in the foreground</th>
</tr>
</thead>
<tbody>
<tr>
<td>• Performance</td>
</tr>
<tr>
<td>• Economic efficiency</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Obstacle to progress</th>
</tr>
</thead>
<tbody>
<tr>
<td>• Demands</td>
</tr>
<tr>
<td>• Blockade</td>
</tr>
</tbody>
</table>

**First Main Category: Essential Foundation**

As part of the first main category (“essential foundation”), all the statements defining ethics as an essential basis for the development of AI-based applications in medicine were summarized.

**Subcategory 1: Awareness**

The first subcategory, “awareness,” highlights the relevance of ethics in development because of the potential dangers and consequences associated with AI:
Because AI is a sharp weapon, [unintelligible] it can be sharpened arbitrarily. But it must be used wisely. And I think one of the biggest difficulties is to anticipate, what does it actually mean when we develop this? [...] this anticipatory ethical question is extremely difficult. [E1; quote A.1]

This subcategory emphasizes the importance of developers being cognizant of the potential uses and challenges that may arise with the subsequent implementation of AI-based applications in medical settings. An additional perspective further reinforces this view:

If we develop something, we always think the application will be used as anticipated in the clinical setting. But we can never be sure, and developers need to be aware of this. [E5; quote A.2]

**Subcategory 2: Consequences**

The second subcategory “consequences,” further emphasizes the importance of ethics in practical development and an associated awareness to prevent consequences such as biases in the data or other potential forms of discrimination from being incorporated into the application:

I think everyone working with AI, especially the field of medicine or [unintelligible], should think of potential consequences involved with it. This does not only include the development teams or companies, but rather anyone. [E4; quote A.3]

Although the previous quote offers a broad view of the ethical considerations in AI for medicine, the next quote from a different expert highlights specific concerns, such as bias and its potential harm to patients:

Yes, well, ethics is super important. [...] Well, when we talk about this bias, when we talk about these false negatives, it’s very important. [...] I am mostly afraid bias. Bias could really harm patients with potentially fatal outcomes. To limit the risk of any bias, we have ongoing discussions in the team. [E5; quote A.4]

**Subcategory 3: Data Protection**

The importance of ethics is also highlighted in terms of the general use of human data in the development of AI-based applications, thereby forming the foundation of the third subcategory:

Well, we actually have this discussion all the time. We at [...] have an ethics working group, for ethical processing and also [unintelligible] and equality. These aspects are always there, especially when you are working with data and people, [unintelligible] data generated by people. [E4; quote A.5]

**Second Main Category: Results in the Foreground**

In the context of the second main category, all statements from the experts are summarized, in which the “Results are in the foreground” of the development of AI-based algorithms.

**Subcategory 1: Performance**

The following quote from the analysis of the third expert interview reflects the result-oriented nature of the development of AI-based applications in medicine, which underlies the formation of the first subcategory:

For me insofar, and I also indirectly deal with it [ethics], but for me it does not represent the first thing. So, if it’s for me, let us say, I want to set up a system first, then it’s also about, I want to set up the system. Ethical aspects do not play a role for me. [...] sounds mean now, but when an IT specialist first trains his models, it’s just about, as banal as it sounds, it’s just about achieving good performance first. [E3; quote B.1]

This result and performance-driven perspective was echoed by another expert, who highlighted the competitive nature of AI development:

But I also believe that there are, let me say, more important things than ethics. Especially with the increased interest in AI, the competition is hard. [...] Developers as well as the applications do need to perform well. [E2; quote B.2]

These statements collectively underscore a tendency within the industry to prioritize performance metrics, which may occasionally overshadow ethical considerations in the drive to advance and remain competitive in the rapidly evolving AI sector.

**Subcategory 2: Economic Efficiency**

The subordinate significance of ethics in performance is also clarified by the following statement in the second subcategory:

I think companies that are in competition, even if they don’t mean it badly, still have the market economic pressure to deliver results, and this can certainly also lead to losing sight of maintaining some ethical boundaries that one would better keep a careful eye on. [E6; quote B.3]

This sentiment is reiterated by another expert who highlights the financial imperatives driving company behavior:

In the end, earning money and making a profit is important to anyone being paid by companies. [...] This might be different in academia, like research, but we all need to focus on creating a product that does financially well, and not trying to be ethically correct. [Interview E2; quote B.4]

These perspectives elucidate the conflict that experts perceive between economic efficiency and ethical conduct in the development of AI-based medical applications.

**Third Main Category: Obstacle to Progress**

The third main category summarizes statements from experts who view ethics as an “obstacle to (technological) progress.”

**Subcategory 1: Demands**

As part of the first subcategory, the “Demands” of ethics are viewed as potential barriers that can stand in the way of AI
technology and the technological progress of AI-based applications in medicine:

I always find it a bit difficult to draw this line between these ethical demands and the limits that then really stand in the way of technology and progress. [E6; quote C.1]

Subcategory 2: Blockade

The perception that ethics can not only hinder current development but also impede future progress in AI forms the basis of the “Blockade” subcategory. This is exemplified by the following statement:

Please stop bothering me on the topic of ethics in AI. It blocks at all corners and edges. [...] Yes, but if I don’t know, how should someone else continue in ten, 20 years so that something comes out of it? [E7; quote C.2]

The aforementioned quote illustrates a dismissive attitude toward ethics as part of the development process of AI-based applications in medicine and thus clarifies the assessment of ethics as an obstacle to (technological) progress. This perspective was reinforced by an additional quote from another expert:

I have no doubt that ethics is important, but it does not help the technological progress of AI. [...] Ethics can really prevent any meaningful advancement. [E6; quote C.3]

Together, these quotes highlight a critical perspective within the AI development community, where ethical concerns, although important, are sometimes seen as obstructions to both immediate technological development and long-term innovation in AI.

Discussion

Principal Findings

The results of the qualitative content analysis revealed a nuanced spectrum of expert opinions regarding the role of ethics in AI development for medical applications. Initially, in the “essential foundation” category, a consensus was observed among experts (e.g., E1 and E5) on the foundational importance of ethics in AI development. This consensus on the foundational role of ethics is based on an understanding of AI’s potential risks and consequences of AI, as exemplified by the anticipatory ethical questions posed by E1 (quote A.1) and the emphasis on uncertainty in application outcomes noted by E5 (quote A.2).

Within the “results in the foreground” category, a shift in perspective becomes apparent. Experts, such as E3 and E2, express views that prioritize performance and competitive outcomes over ethical considerations (quotes B.1 and B.2). This shift suggests a conflict between ethical integrity and market-driven objectives, with the latter often taking precedence in the fast-paced competitive landscape of AI development.

In the “obstacle to progress” category, the tension between ethical demands and technological advancement is further articulated. Expert E6, for instance, acknowledged the difficulty of reconciling ethical demands with the limits imposed on technology and progress (quote C.1). This sentiment is echoed by expert E7, who expresses frustration with ethics perceived as a blockade of development (quote C.2). These perspectives underscore a critical view within the AI development community, where ethical concerns, although recognized as important, are sometimes seen as obstacles to immediate technological development and long-term innovation.

This variety of opinions, ranging from viewing ethics as foundational to considering them as impediments, reflects the complex and multifaceted nature of AI development in medicine. This demonstrates that although there is a general recognition of the importance of ethics, the extent to which it is prioritized differs significantly among experts. This diversity highlights the challenges in balancing ethical considerations with other developmental goals, such as performance optimization, economic viability, and technological innovation.

The analysis of the expert interviews identified 3 critical themes: first, the incompleteness of data and the far-reaching consequences associated with it; second, the renunciation of ethical requirements because of economic pressure; and third, the opinion that adhering to ethical standards would stand in the way of technological progress. These themes, reflecting a spectrum of perspectives from foundational importance to perceived obstacles, are explored in detail in subsequent sections, providing a deeper understanding of the multifaceted nature of ethics in AI development for medicine.

Incompleteness of Data

Quote A.4 (section Subcategory 2: Consequences) refers to the relevance of biases in the data. The lack of representativeness of the data, which underlies the development of AI-based applications, has been cited as a fundamental potential bias. Although awareness of the potential consequences, such as discrimination against certain population groups, is a crucial first step, it is not enough to merely recognize the issue to avoid potentially significant consequences [38]. Therefore, active measures must be taken to prevent these biases and ensure that AI-based applications do not perpetuate or exacerbate inequalities, thereby limiting potential harm.

To mitigate bias risks, developers should adopt comprehensive strategies, such as inclusive data collection methods, algorithmic audits, thorough testing across various demographic groups, and ongoing bias monitoring throughout the AI application lifecycle. As highlighted in quote A.1, the anticipatory ethical question in AI development is “extremely difficult,” underscoring the complexity of ensuring that AI systems are ethically sound and free from biases that could lead to discrimination or harm. Interdisciplinary teams, including ethicists and representatives from diverse communities, should guide the development process to ensure that ethical considerations are at the forefront of AI development.

A potential consequence of nonrepresentative data, as highlighted in quote A.4, includes “false negatives” in medicine, which are test results that incorrectly turn out to be negative despite the presence of diagnostic features of the disease under investigation [25]. However, it is also critical to recognize that
the same issue of nonrepresentativeness can lead to “false positives,” where tests incorrectly indicate the presence of a condition that is actually absent [25]. Both types of diagnostic inaccuracies have serious implications for patient care and treatment outcomes. This is further compounded by the sentiment expressed in quote A.3, where the need for everyone working with AI, especially in medicine, to consider the potential consequences of their work is emphasized, indicating a broader responsibility beyond development teams. This emphasizes the need for a comprehensive approach to diagnostic accuracy that accounts for both the presence of representative data and various factors influencing AI performance, extending beyond data representativeness [26]. Accuracy is also determined by the quality and variety of information subject to analysis from AI-based applications, including clinical, laboratory, and patient-reported data [39]. Furthermore, how AI processes and interprets this information, such as through its underlying algorithms and decision-making logic, is highly important for diagnostic accuracy [40]. There must be a match between the design purpose of the algorithm and real-world scenarios in which it is applied.

Moreover, the diagnostic accuracy of AI-based applications depends substantially on the proficiency with which health care professionals use these tools and their capacity to interpret and act on AI-generated recommendations [41]. For instance, if AI applications are used beyond their original scope without proper recalibration or validation for new populations or diseases, there is a risk of introducing errors, including false negatives and false positives [25].

False negatives in a clinical context can lead to physicians feeling a false sense of security and the diseases of patients remaining untreated for a long time [25]. Conversely, false positives can result in unnecessary treatments when a test erroneously indicates the presence of a disease, leading to significant consequences, such as unwarranted radiation exposure [25]. The psychological impact on patients, resulting from both false negatives and false positives, is a further concern that merits attention because of its effect on patient well-being and trust in medical systems.

The ethical implications of AI development, particularly when personal data are used, are highlighted in quote A.5 (section Subcategory 3: Data Protection). The use of training data for diagnosing specific diseases requires a careful ethical approach, particularly to understand the personal and clinical contexts from which such data are derived. This is particularly important for diseases that restrict the ability of the affected individuals to provide informed consent. Furthermore, ongoing discussions within ethics working groups about ethical processing, as mentioned in quote A.5, play a crucial role in safeguarding the dignity and rights of individuals whose data are used in these systems. Therefore, developers must recognize the sensitivity of medical data and the need for ethical considerations to be integrated from the outset of AI development for medical applications. Such early integration of ethics serves not only to enhance the accuracy and reliability of AI tools but also to safeguard the dignity and rights of individuals whose data are used in these systems.

**Economic Pressure**

The quotes from the second main category “results in the foreground” suggest that although the interviewed experts are aware of the relevance of ethics in the development of AI-based applications, it is in conflict with their own or demanded result orientation. A possible reason for the experts’ assessment is mentioned in quote B.3 (section Subcategory 2: Economic Efficiency). The profitability of AI developing companies is cited as one of the reasons why ethics is subordinate to the results in practice. Companies’ economic success pressure is decisive for the success pressure of all the employees involved in development. This conflict is further illustrated in quote B.2, where an expert highlights the competitive nature of AI development, suggesting that there are “more important things than ethics” in the context of existing competition. This perspective underscores the challenge of balancing ethical considerations with the need for AI applications to perform well in competitive markets.

As quote B.1 (section Subcategory 1: Performance) illustrates, the best possible performance is the focus of the development. Ethics indirectly plays a role here; quote B.3 implies, in this sense, the possibility of crossing “ethical boundaries” in favor of profitability. In addition to the deliberate crossing of boundaries, this statement also implies the possibility of unconscious disregard for ethics in the development of AI-based applications. The subordinate role of ethics in profitability in development and the associated noncompliance with potential boundaries is particularly severe, as the field of application is medicine. The sentiment of economic pressure overshadowing ethical considerations is also echoed in quote B.4, in which an expert states the importance of focusing on creating a product that is financially well, often at the expense of being ethically correct.

In addition to the relevance of ethics in relation to the use of human data and the potential consequences of a lack of representativeness, patient safety should always be at the center of the development of medical products and technologies. An excessive focus on the profitability of an application can lead to the marketing of immature or faulty products, which threaten patient welfare. Furthermore, as highlighted in quote B.3, the pursuit of profitability can sometimes lead developers to overlooking ethical boundaries, potentially resulting in products that have not been thoroughly evaluated for ethical considerations and patient safety. In addition to a direct threat to patient welfare and safety, a high susceptibility to error can also lead to rejection by users and a potentially irretrievable loss of trust [42].

**Obstacle to Progress**

Although the second main category cites result orientation because of economic pressure as a reason for the subordination of ethics, the third main category summarizes statements that view ethics as an “obstacle to progress.” The statements of experts in this category clearly show a rejection of ethics because of various demands and boundaries that are perceived as obstacles to the development of AI-based applications. Although no specific reasons for this assessment are provided, based on the knowledge of the steps relevant to development,
Consequences of Neglecting Ethics in the Development of AI-Based Applications in Medicine

Overview

If ethics is not considered in the development process of AI-based applications, it can have far-reaching consequences for patients and physicians, such as loss of trust and erosion of patient-centered care. This section focuses on the possible consequences of neglecting ethics when developing AI-based applications in medicine. In this context, the consequences for patients and likely main users (physicians) were considered.

Possible Consequences for Patients

If those responsible do not consider or only marginally consider the basic ethical principles in the development process of AI-based applications, various indirect and direct consequences can occur for the patients in whom the respective AI-based applications are used. The following examples illustrate the possible consequences of not considering ethical principles in the development process of AI in medicine:

- Misdiagnosis and diminished therapy outcomes: a lack of ethical considerations in the practical development process of AI-based applications can lead to biases in the training data used for development. For example, if the applications are used for diagnosis, the lack of representativeness of the data for certain population groups or individuals can lead to a higher susceptibility to errors. The results presented by AI can lead to potentially significant consequences for patients, such as overtreatment or undertreatment, resulting in diminished therapeutic outcomes, particularly in the absence of control by users [11]. These errors, stemming from a misdiagnosis because of unrepresentative data, challenge the principle of justice by threatening equitable medical care and contravene the principle of nonmaleficence by risking patient harm through inappropriate medical procedures [24]. Moreover, susceptibility to errors may directly compromise patient outcomes, especially when undertreatment occurs because of delayed or missed treatments from false-negative results [16]. The interrelated consequences of misdiagnosis and therapy outcomes highlight the critical need for user oversight and inclusion of diverse data sets in AI development to uphold ethical standards and patient care quality.

- Loss of trust: faulty diagnoses and the possibility of AI-based applications yielding discriminatory results can significantly undermine patient trust [44]. Such erosion of trust may lead patients to view AI-based medical applications skeptically, potentially refraining from using them in their treatment. This skepticism can hinder the integration of advanced AI tools in health care, which, if more accurate than physicians' assessments, could otherwise enhance patient outcomes. A loss of trust not only impedes technological adoption but can also indirectly challenge the principle of care, which is dedicated to optimizing patient welfare. Furthermore, patient reluctance to embrace AI solutions may inadvertently perpetuate inequalities in health care, particularly if AI facilitates more effective and efficient clinical practice. The reluctance to use AI technologies could result in disparity in care quality, as

it can be assumed that the statements primarily refer to regulations and requirements in the sense of a necessary positive vote by ethics committees. For data collection, use, or evaluation in the context of developing AI-based applications, compliance with certain boundaries and regulations is indispensable, not only in the medical context. However, this essential compliance is sometimes perceived by experts as a balancing act, where meeting ethical demands can create challenges in advancing AI technology (quote C.1).

These boundaries and regulations serve to protect the participants and their data. If patient data are to be used, a positive vote from an ethics committee that certifies the safety of patients and their data is necessary to begin with the respective research and data use. As ethics committees’ decisions can be time intensive depending on the type of planned research or data use and often require corrections on the part of the applicants, it is assumed that the necessity of a positive vote is one of the reasons that is viewed as an obstacle to progress. Furthermore, as highlighted in quote C.2, frustration with ethics being viewed as a blockade is evident: “Please stop bothering me on the topic of ethics in AI. It blocks at all corners and edges,” illustrating the tension between the desire for rapid AI development and the need for ethical oversight. Although it can be assumed that AI-based applications would be developed faster if no vote from an ethics committee was necessary and patient data could be used directly, the resulting consequences for patients and citizens (think of the insurance industry) at least require critical evaluation.

Furthermore, although the need for a positive vote by an ethics committee can be anticipated as a perceived obstacle to progress in the development of AI by experts, it is also important to consider ongoing regulatory efforts, such as the proposed “Artificial Intelligence Act” by the European Parliament [43]. This regulation aims to harmonize rules on AI across the European Union, focusing on human-centric and trustworthy AI. The Act emphasizes the protection of health, safety, fundamental rights, and environmental concerns from potential harm caused by AI systems. It includes specific recommendations for high-risk AI systems, such as AI-based applications for medicine, demanding transparency, accountability, and accuracy in AI applications, especially those that may significantly impact individuals’ rights and safety. The Act further acknowledges the ethical considerations in AI development and underscores the need for AI systems to adhere to robust ethical and legal standards. The regulatory requirement to adhere to ethical standards, as mandated by the Act, could further reinforce the perception of ethics and regulations being an obstacle, highlighting the tension between rapid technological advancement and the need for responsible innovation. In addition, quote C.3 conveys a sentiment shared by some experts that although ethical considerations are undeniably important, they are sometimes viewed as hindrances to meaningful AI advancement, further highlighting the complex dynamics between ethical considerations and the pursuit of technological progress in AI.
physicians may be limited in their capabilities without AI support, ultimately affecting the standard of care provided. Moreover, an erosion or lack of trust in AI because of missing ethical oversight in development could extend to the physician-patient relationship and the overall health care sector. Moreover, an erosion or lack of trust in AI because of missing ethical oversight in development could extend to the physician-patient relationship and the overall health care sector [45]. This could lead to a general skepticism toward medical advice and a hesitation to participate in newer forms of treatment, potentially reverting to more traditional but less efficient methods. The physician-patient relationship is foundational to effective health care, as it relies on mutual trust and the belief that the best possible treatment options are being used, including ethically developed AI applications.

- Data misuse: a lack of consideration of ethics in the development of AI-based applications can lead to violations of existing data protection laws and misuse of patient data [46]. Patients who provide their data for research purposes and for the development of new applications in medicine must be able to rely on careful and legally compliant handling of their data, particularly in terms of informed consent and cybersecurity. Given the lack of traceability, informed consent is crucial, as patients must have a clear understanding of how their data will be used and the ability to consent to specific uses. This is of particular importance because health-related data include personal and sensitive information about patients. Ignoring existing regulations and ethical principles can result in highly sensitive patient data becoming accessible to companies, organizations, or individuals without consent [46]. This could have far-reaching consequences such as compromising patient privacy, enabling identity theft, or even affecting the broader integrity of medical research and public trust in the health care system. Similarly, robust cybersecurity measures are essential to protect sensitive health information from unauthorized access and breaches. Failure to implement such measures can lead to the exposure of personal health data, resulting in a loss of patient trust, potential harm, and a violation of the autonomy of patients if they lose control over their own data.

- Erosion of patient-centered care: the exclusion of patient values and preferences during the development of AI-based medical applications can have profound consequences. When AI systems are designed without a thorough understanding of patient autonomy, self-determination, and individual health goals, there is a risk of eroding the essence of patient-centered care [47]. AI recommendations that do not account for these personal factors might lead to a mechanical and less social approach to health care that could disregard the nuanced needs and desires of patients. For example, if AI tools are optimized solely for clinical efficiency without considering patient comfort and personal treatment preferences, they may suggest interventions that patients find unacceptable or intrusive. This misalignment can result in decreased adherence to treatment plans, loss of trust in the physician-patient relationship, and diminished health outcomes [48]. Given the importance of autonomy in the physician-patient relationship and patient care in general, AI-based applications should be designed to support a shared decision-making model in which AI assists the therapeutic process rather than diminishing it. This would ensure that AI acts as an aid rather than a replacement for the human element in health care, empowering patients to be active participants in their treatment decisions rather than passive recipients of care.

**Potential Consequences for Physicians**

In addition to the significant consequences for patients, the lack of ethical consideration in the development process of AI-based applications in medicine can also lead to equally relevant impacts on anticipated primary users of the technology. Although the following examples primarily aim to illustrate the direct consequences for physicians, they also indirectly affect the patients being treated:

- Loss of credibility: potential errors in diagnosis or treatment recommendations resulting from inadequately trained AI applications can also significantly influence the societal image of the medical profession and its associated credibility [49]. Assuming that physicians continue to serve as the link between technology and patients, erroneous decisions based on the use of AI in medicine can be directly associated with the decision-making abilities of physicians, which can negatively impact their credibility and trust in the medical community [44]. Knowledge about the potential for discrimination of certain population groups by AI-based applications, which do not consider ethical guidelines in their development, can further shake patients’ beliefs that physicians guarantee equal treatment for all. Because a patient’s medical treatment often appears nontransparent and incomprehensible, the credibility of the medical community is an essential prerequisite for the physician-patient relationship [49].

- Rejection: the lack of consideration of ethics in the development of AI-based applications for use in a clinical context can lead to both indirect (eg, because of the consequences of incorrect diagnoses) and direct (eg, because of the lack of consideration of ethical principles) rejection of the technology by physicians. The rejection of AI-based applications can also significantly influence the societal and technological progress in medicine. Without the acceptance and trust of prospective primary users of the technology, the widespread use of AI-based applications in medicine is unlikely, as economic incentives for development are lacking. A rejecting attitude on the part of physicians can in this context also negatively impact future medical care quality considering the expected advantages of using AI in medicine [46].

- Legal consequences: the use of AI-based applications developed without considering ethical principles can lead to various legal consequences for users [50]. In addition to consequences based on state legislation and jurisprudence, professional legal consequences for physicians are also conceivable when using AI-based applications without considering ethical principles, as they form the basis of medical action. Besides the direct legal implications for physicians, health care organizations, such as hospitals,
clinics, or research institutions, may also be subject to significant responsibilities and potential liabilities when deploying AI-based applications that may not fully align with ethical and regulatory standards. In the case of erroneous AI decisions, which directly or indirectly result in diminished patient outcomes, the question of legal liability often remains unanswered [51]. As AI-based applications in medicine are likely to continue to be used and developed in a supportive role, it is assumed that the final decision-making and treatment recommendations will remain the responsibility of physicians. Thus, physicians not only act as a link between technologies and patients but also play a central role in adhering to ethical principles in medical care. Against this background, the use of AI-based applications in medicine developed without considering ethics can have legal consequences for both developers and users. In addition to the legal consequences of erroneous medical treatments, the use of AI-based applications without considering ethical principles also raises questions regarding the liability for violation of existing data protection and equal treatment laws [51]. In particular, failure to comply with data protection laws can compound these legal issues. Violations of patients’ privacy rights through the mishandling of sensitive patient data, whether because of inadequate security measures, hacks, or unauthorized data sharing, may subject various entities, such as hospitals, clinics, research institutions, AI technology developers, and users to significant legal liability [50]. These data breaches not only compromise patient confidentiality but also could lead to a risk of regulatory sanctions for the involved entities, including substantial fines and potentially the loss of professional licenses. Therefore, AI development processes should incorporate robust data protection protocols to prevent legal repercussions and consequences for both patients and physicians. Adherence to ethical and legal standards should not merely be a regulatory requirement but a fundamental component of responsible and trustworthy health care innovation, vital for maintaining the integrity of patient care and the broader medical profession.

Limitations
This study’s exploration of expert perspectives on ethics in AI development for medical applications, although insightful, encounters several limitations that are important to acknowledge. First, the geographical focus of the study was confined to Germany, potentially limiting the applicability of its findings to a global context in which cultural, legal, and ethical norms may vary. The selection of experts, although experienced in the development of AI-based applications in medicine, represents a relatively small and specific segment of the broader field. Moreover, the focus of the study, predominantly on experts with technical backgrounds in the development of AI-based applications, may lead to a narrowed perspective, given the lack of input from ethical professionals. Furthermore, the subjective nature of expert interviews should be considered because the responses are influenced by each expert’s personal experiences and potential biases, which may not comprehensively represent the spectrum of views in the field.

Methodologically, the study’s qualitative approach and reliance on secondary analysis of expert interviews inherently limits the generalizability of the results. Interpretations may be influenced by the research team’s perspectives, and certain nuances in experts’ statements may be overlooked. Although this study presents a secondary analysis of existing data, it is important to recognize the possibility of confirmation and selection bias during the initial data collection phase. The research methodology used could have unintentionally emphasized certain themes or perspectives, potentially aligning with the original researchers’ preconceived notions or expectations. In addition, because of the limited number of experts included in the analysis and incomplete data saturation in some subcategories, certain aspects may not have been fully explored.

Furthermore, the findings of this study reflect a specific point in time in a rapidly evolving field. Therefore, the perspectives and opinions of experts may change as new developments, regulations, and ethical guidelines emerge. Although substantial, the focus on the development of AI-based applications in medicine does not encompass the entire spectrum of AI applications within the health care sector, excluding administrative and operational uses. Language and translation limitations may also have affected the study, as the original German interviews were translated into the English language. The subtle nuances of language and cultural context might be lost or misinterpreted in this translation process.

To address these limitations and enrich future research in this area, it is recommended that subsequent studies incorporate a broader and more diverse pool of experts, including professionals from ethical, legal, and patient advocacy backgrounds. Expanding the geographical scope to include experts from various cultural and legal contexts would also provide a global perspective on the ethical implications of developing AI-based applications for medicine. Methodologically, integrating both qualitative and quantitative approaches could offer a more comprehensive view, although ongoing research is required, considering the rapid advancements in AI and evolving ethical standards. By expanding the scope and methodology of future studies, a more nuanced and representative exploration of the ethical landscape of AI development for medical applications can be achieved.

Summary and Outlook
This study explored the importance of ethics in the development of AI-based medical applications by analyzing interviews with experts in the field of AI development. There was substantial variance in the assessment of the importance of ethics in the development of the AI-based applications. Although some of the interviewed experts classified ethics as an essential basis for development, others focused on good performance or economic efficiency. The results of the qualitative analysis also suggest that ethics is seen by some experts as an obstacle to progress, implying that it will be given little importance in the further development of AI-based applications. In addition to the subsequent discussion of the content analysis results, a particular focus was placed on the consequences that could arise from the lack of ethical considerations in the development of AI-based applications in medicine.
Although the results do not allow for generalization, because of the number of interviewees and the selected qualitative research method not meeting representative demands, the statements of the interviewed experts should be seen as an essential basis for further research and discussions because of recurring motives and new insights. A lack of ethical considerations in the development of AI-based applications can have significant consequences for patients. In addition to the danger of misconduct (e.g., because of a lack of representativeness of the data sets used for development), a lack of consideration of ethical principles in the development of AI-based applications can also lead to a loss of trust from patients and potentially diminished therapy outcomes. When considering the possible impacts on physicians, the lack of consideration of ethics in the development process can lead to loss of credibility and rejection of technology.

Owing to technological progress in the field of AI, further reinforced, for example, by the development and broad availability of AI-based chat applications such as ChatGPT, there has been ongoing effort to develop guidelines and laws to guide the development and use of AI. Although such regulatory efforts, such as the “Artificial Intelligence Act” for harmonized rules on AI from the European Parliament, aim to provide a comprehensive regulatory framework and guideline for the development and use of AI, there is ongoing criticism and discussion about the adequacy and effectiveness of these guidelines in the rapidly evolving field of AI. In this context, it is important to emphasize that the sole availability of guidelines and laws does not ensure compliance. Therefore, although guidelines and laws are important to guide the development and use of AI, especially in the field of medicine, and when dealing with sensitive patient data, more work needs to be done to ensure compliance.

Moreover, the question arises as to whether mere adherence to these guidelines and laws is sufficient for the development of ethical AI. Guidelines often provide a baseline for legal compliance, but ethical AI development demands a deeper and more nuanced understanding and application of ethical principles. Ethical AI goes beyond legal requirements to encompass ethical principles, such as respect for autonomy or justice in its algorithms, data handling, and decision-making processes. This requires continuous ethical assessment and reflection throughout the lifecycle of AI-based applications, from development to deployment, and beyond. Consequently, although following established guidelines is an important step in the development of AI, it is not the endpoint. Developers and users of AI-based applications in medicine need to engage in an ongoing dialog with diverse stakeholders such as ethicists, patients, and the broader community to anticipate, identify, and address emerging ethical challenges. This approach ensures that the development of AI is not just about complying with regulations but is intrinsically driven by a commitment to ethical responsibility and the betterment of patient care.

Furthermore, possible reasons for noncompliance with potential guidelines and low prioritization of ethics, such as the need for economic efficiency, should be critically examined. This includes assessing perspectives that view ethics as an obstacle to progress, as noted by some participating experts. Such critical evaluation is vital for ensuring the ethical development of AI-based applications, particularly in the field of medicine. Ethical considerations are fundamental to every approval process for AI-based applications to ensure the best possible and equal medical care for patients. Therefore, physicians should critically question the use of AI-based applications in the clinical context. In this regard, there needs to be a sufficient availability of opportunities to acquire further competencies to promote an understanding of technology and the related relevance of ethics. Only in this manner can the safety and best possible treatment of patients be ensured, as well as medical and technological progress, through AI.

Conflicts of Interest
None declared.

References
2. Eysenbach G. The role of ChatGPT, generative language models, and artificial intelligence in medical education: a conversation with ChatGPT and a call for papers. JMIR Med Educ 2023 Mar 06;9:e46885 [FREE Full text] [doi: 10.2196/46885] [Medline: 36863937]
34. Weidener L, Fischer M. Artificial intelligence teaching as part of medical education: qualitative analysis of expert interviews. JMIR Med Educ 2023 Apr 24;9:e46428 [FREE Full text] [doi: 10.2196/46428] [Medline: 36946094]


Abbreviations

AI: artificial intelligence
CDSS: Clinical Decision Support Systems
ML: machine learning
Identifying Symptoms Prior to Pancreatic Ductal Adenocarcinoma Diagnosis in Real-World Care Settings: Natural Language Processing Approach

Fagen Xie1, PhD; Jenny Chang1, MPH; Tiffany Luong1, MPH; Bechien Wu1, MD, MPH; Eva Lustigova1, MPH; Eva Shrader2, MS; Wansu Chen1, PhD

1Department of Research and Evaluation, Kaiser Permanente Southern California, Pasadena, CA, United States
2Pancreatic Cancer Action Network, Manhattan Beach, CA, United States

Corresponding Author:
Fagen Xie, PhD
Department of Research and Evaluation
Kaiser Permanente Southern California
100 S Los Robles Avenue
Pasadena, CA, 91101
United States
Phone: 1 6265643294
Email: fagen.xie@kp.org

Abstract

Background: Pancreatic cancer is the third leading cause of cancer deaths in the United States. Pancreatic ductal adenocarcinoma (PDAC) is the most common form of pancreatic cancer, accounting for up to 90% of all cases. Patient-reported symptoms are often the triggers of cancer diagnosis and therefore, understanding the PDAC-associated symptoms and the timing of symptom onset could facilitate early detection of PDAC.

Objective: This paper aims to develop a natural language processing (NLP) algorithm to capture symptoms associated with PDAC from clinical notes within a large integrated health care system.

Methods: We used unstructured data within 2 years prior to PDAC diagnosis between 2010 and 2019 and among matched patients without PDAC to identify 17 PDAC-related symptoms. Related terms and phrases were first compiled from publicly available resources and then recursively reviewed and enriched with input from clinicians and chart review. A computerized NLP algorithm was iteratively developed and fine-trained via multiple rounds of chart review followed by adjudication. Finally, the developed algorithm was applied to the validation data set to assess performance and to the study implementation notes.

Results: A total of 408,147 and 709,789 notes were retrieved from 2611 patients with PDAC and 10,085 matched patients without PDAC, respectively. In descending order, the symptom distribution of the study implementation notes ranged from 4.98% for abdominal or epigastric pain to 0.05% for upper extremity deep vein thrombosis in the PDAC group, and from 1.75% for back pain to 0.01% for pale stool in the non-PDAC group. Validation of the NLP algorithm against adjudicated chart review results of 1000 notes showed that precision ranged from 98.9% (jaundice) to 84% (upper extremity deep vein thrombosis), recall ranged from 98.1% (weight loss) to 82.8% (epigastric bloating), and F1-scores ranged from 0.97 (jaundice) to 0.86 (depression).

Conclusions: The developed and validated NLP algorithm could be used for the early detection of PDAC.

(JMIR AI 2024;3:e51240) doi:10.2196/51240

KEYWORDS

cancer; pancreatic ductal adenocarcinoma; symptom; clinical note; electronic health record; natural language processing; computerized algorithm; pancreatic cancer; cancer death; abdominal pain; pain; validation; detection; pancreas

Introduction

Pancreatic cancer is the third leading cause of cancer deaths in the United States, with 50,550 estimated deaths in 2023 [1]. Pancreatic ductal adenocarcinoma (PDAC), which accounts for 90% of pancreatic cancer cases, is the most common form of pancreatic cancer. The age- and sex-adjusted incidence has continued to increase, reaching 13.3 per 100,000 in 2015-2019,
and the overall 5-year survival remains poor at only 12.5% [2]. Despite technological advances, diagnosis of pancreatic cancer remains very late, with more than 50% of patients having distant metastases at the time of diagnosis [2-4].

Patient-reported symptoms are often the trigger for evaluation that eventually leads to a diagnosis of pancreatic cancer [5,6]. The reported prevalence of symptoms associated with PDAC has largely varied due to many factors, such as study design and data sources [6-10]. Additionally, previously published studies have been based on patient surveys [6,7] or structured electronic health records (EHRs) [8-10]. However, structured data can be inaccurate [11,12] and incomplete [13], especially for signs and symptoms. On the other hand, signs and symptoms are frequently collected and documented in the clinical notes by care providers via free text within the EHRs. Therefore, extracting signs and symptoms from clinical notes offers a key opportunity for the early detection of pancreatic cancer, which can lead to more timely interventions that improve survival.

Identification of PDAC-related symptoms from clinical notes based on EHRs is a challenge because signs or symptoms are typically not well-documented in a structured format within an EHR system, and specific techniques are required for data processing and analysis. Natural language processing (NLP), a field of computer-based methods aimed at standardizing and analyzing free text, processes unstructured data through information extraction from natural language and semantic representation learning for information retrieval, classifications, and predictions [14]. Numerous innovative NLP applications have been developed across various clinical domains in support of medical research, public health surveillance, clinical decision making, and outcome predictions [15-19]. Early NLP applications have largely focused on rule-based approaches [15,16], while recent NLP applications utilize state-of-the-art machine learning [17] or deep learning approaches via transformer learning models [18-20]. Rule-based NLP techniques have been widely used to extract signs and symptoms from free-text narratives in past years [21-26]. To the best of our knowledge, we are not aware of previous studies systematically analyzing pancreatic cancer–related symptoms from clinical notes via NLP. The purpose of this study is to develop and validate a comprehensive NLP algorithm and process to effectively identify PDAC-related symptoms prior to diagnosis within a large integrated health system.

**Methods**

**Study Setting**

Kaiser Permanente Southern California (KPSC) is an integrated health care system providing comprehensive medical services to over 4.8 million members across 15 large medical centers and more than 250 medical offices throughout the Southern California region. The demographic characteristics of KPSC members are diverse and largely representative of the residents in Southern California [27]. Members obtain their health insurance through group plans, individual plans, and Medicare and Medicaid programs and represent >260 ethnicities and >150 spoken languages. KPSC’s extensive EHR data contains individual-level structured data (ie, diagnosis codes, procedure codes, medications, immunization records, laboratory results, and pregnancy episodes and outcomes) and unstructured data (ie, free-text clinical notes, radiology reports, pathology reports, imaging, and videos). KPSC’s EHR covers all medical visits across all health care settings (eg, outpatient, inpatient, and emergency department). Clinical care of KPSC members provided by external contracted providers is captured in the EHR through reimbursement claim requests.

**Ethical Considerations**

The study protocol was reviewed and approved by the KPSC Institutional Review Board (approval no. 12849) with a waiver of the requirement for informed consent.

**Study Population Identification**

This study was a nested case-control study of KPSC patients aged 18-84 years between 2010 and 2019. Patients diagnosed with PDAC were identified through KPSC’s cancer registry. Patients with a history of acute or chronic pancreatitis, without a clinic-based visit within 3 to 24 months prior to the diagnosis, with chemotherapy or infusion treatment, or with less than 20 months of health plan enrollment or pregnancy within 2 years prior to the diagnosis date were excluded. Among the patients with PDAC, the date of diagnosis was defined as the index date. For each PDAC case, up to 4 controls were selected from a group of patients without PDAC on the index date of the matched cases. Controls could develop PDAC 1 year after the index date. The above study criteria identified a total of 2611 eligible patients with PDAC and 10,085 corresponding matched patients without PDAC during the study period. The study participant identification and NLP process is shown in Figure 1.
PDAC Symptom Selection
We initially identified 24 PDAC-related symptoms based on literature reviews and clinicians’ input. A survey was conducted among the Consortium for the Study of Pancreatitis, Diabetes, and Pancreatic Cancer working group members [28] to determine the relative importance of the 24 potential symptoms. Based on the ranking of importance, a total of 17 symptoms were finally selected. In this study, we considered abdominal pain and epigastric pain as a combined symptom (abdominal or epigastric pain) and anorexia and early satiety as a combined symptom of (anorexia or early satiety) due to the difficulty of distinguishing them in clinical notes or patient-provider communications. The deep vein thrombosis (DVT) symptom was included in our study because DVT risk is high in patients with pancreatic cancer [29], and the symptom was further delineated into upper and lower DVT.

PDAC Symptom Keyword Selection
First, we compiled a list of phrases or terms relevant to the 17 symptoms based on previous literature [21-23] or symptom ontologies in the Unified Medical Language System [30]. The list was then reviewed and enriched by the experienced study gastroenterologist and enhanced by manual data annotation processing (refer to “Data Annotation” subsection for details). In addition, we used a word embedding model, Word2vec [31,32], to capture possible relevant phrases and terms, including misspelled terms, for each symptom. The compiled comprehensive phrases and terms for these 17 symptoms are summarized in Table S1 in Multimedia Appendix 1. The PDAC symptoms can be determined by a single phrase or term except for the DVT symptom. The DVT symptom was determined by 3 sets of terms, which included location (eg, leg or arm), feeling or appearance (eg, pain or swollen), and laterality (eg, left or right), rather than a single phrase or term.

Extraction and Preprocessing of Study Notes
Clinical notes and patient communication messages (telephone or email) within 2 years prior to the index date of PDAC cases and their matched controls (referred to as “notes” hereafter) were extracted from the KPSC EHR system. Notes associated with certain medical encounters (eg, surgery), note types (eg, patient instructions or anesthesia), and department specialties (eg, health education) were excluded from the analysis because symptoms of interest were unlikely to be present in these notes (Table S2 in Multimedia Appendix 1). The extracted notes were then preprocessed through the following steps: (1) lowercase conversion, sentence splitting, and word tokenization [33]; (2) removal of nondigital or nonletter characters except for spaces, periods, commas, question marks, colons, and semicolons; (3) standardization of abbreviated words; and (4) correction of misspelled words based on the Word2vec model supplemented by an internal spelling correction file developed in previous studies [23,25].

Training, Validation, and Implementation Data Sets
Our study involved 2 phases of training and validation. The first phase used the notes of 100 randomly selected PDAC cases. The second phase used a subset of notes from both PDAC cases and controls. Details of the sample selection for training and validation are summarized in Table S3 in Multimedia Appendix 1. Notes that were not used for training or validation formed the study implementation data set.

Data Annotation
Notes from both the training and validation data sets were manually reviewed by trained research annotators to indicate the presence of the 17 symptoms based on the established terms and phrases (Table S1 in Multimedia Appendix 1) and inclusion and exclusion criteria (Table S4 in Multimedia Appendix 1). The note annotation process was based on a computer-assisted approach. First, notes from the training and validation data sets were exported into a spreadsheet and the prespecified terms (Table S1 in Multimedia Appendix 1) were highlighted. Second, for each note, the annotators reviewed the notes to label the presence of each of the 17 symptoms. Third, any ambiguous notes were fully discussed during weekly study team meetings until a consensus was reached. Cases that were difficult to determine were reported to the study gastroenterologist for adjudication.

A subset of the training data set in the first phase (n=2795 notes) was double-reviewed (ie, 2 annotators independently reviewed the same set of notes). The results from the 2 annotators were

---

**Figure 1.** Schematic diagram of the NLP algorithm to identify the pancreatic ductal adenocarcinoma–related symptoms. EHR: electronic health record; NLP: natural language processing.
compared and inconsistencies between them were discussed until a consensus was reached. If the annotators did not reach a consensus, the note was reviewed and adjudicated by the study gastroenterologist.

Finally, the adjudicated results were documented as the gold standard for training and validation of the NLP algorithm.

NLP Algorithm Development

Algorithm development involved 2 phases of training. For each phase, we used the annotated training data set to develop or refine a rule-based computerized algorithm via an iterative process to determine the presence of the 17 symptoms in each note. First, the notes were analyzed based on the phrase or terms and patterns that indicated the presence or absence of each symptom (Table S1 in Multimedia Appendix 1). The algorithm was then processed to search for patterns of inclusion or exclusion to determine the status of each symptom (Table S4 in Multimedia Appendix 1). A list of negated terms (eg, “ruled out” or “negative for”), uncertain or probable terms (eg, “presumably”), definite terms (eg, “positive for”), history terms (eg, “several years ago”), non-patient person terms (eg, referring to a family member), and general description terms (eg, “please return to ED if you have any of the following symptoms”) were compiled from the training data sets. The compiled terms were enriched via the repeated test-revise strategy against the chart review results within each training subset until the algorithm performance reached an acceptable threshold (ie, positive predictive value [PPV]≥90%). The discordant cases between the algorithm and manually annotated results for each subset were further reviewed and adjudicated among the annotators and study team until a consensus was reached.

Specifically, each symptom for each note was first determined at the sentence level based on the following criteria:

1. A sentence defaulted as “no” if any exclusion criterion in Table S4 in Multimedia Appendix 1 was met.

2. The symptom was considered absent if the sentence met any of the following situations:
   - The sentence did not contain any defined terms listed in Table S1 in Multimedia Appendix 1.
   - The negated description was associated with defined terms listed in Table S1 in Multimedia Appendix 1. Examples included “patient denied vomiting/nausea,” “ruled out jaundice,” and “no pruritus.”
   - The description of the symptom did not refer to an actual situation. For example, “return if you experience epigastric bloating” and “glipizide side effects including loss of appetite, nausea, vomiting, weight gain.”
   - A probable or uncertain description was associated with the symptom. For example, “patient with anxiety and likely depression” and “patient informed that there may be pruritis or pain.”
   - The symptoms were associated with a historical term or date relative to the clinical note date. For example, “patient had abdominal pain two years ago” and “patient had jaundice in 2007.”
   - The symptom description was related to family history, such as “family history: mother anxiety” and “patient family history: daughter with depression.”

3. A symptom was classified as “yes” for any of the following situations:
   - The symptom contained a symptom of interest and the symptom was marked as “yes,” “x,” or “+”. A symptom was classified as “yes” if the response to a symptom question was affirmative or if the symptom was marked on the symptom list.
   - The symptom was listed under the diagnosis section (except for DVT), chief complaint section, symptom section, and history of present illness section of the clinical note. For example, “chief complaint: abdominal pain,” “primary encounter diagnosis anxiety disorder,” and “jaundice 782.4.”
   - The symptom was described as treated or indicated by medication within nonhospitalization encounters.
   - The symptom was documented or reported to be present at the time of visit or messaging. For example, “pt complaint of 55 lb weight loss since March 2009” and “patient here for several weeks of abdominal pain.”
   - The sentence contained a definite term associated with a symptom of interest. Examples included “positive for fatigue and weight loss,” “patient reports anorexia,” and “patient presents with anxiety, depression, insomnia.”

4. The sentence-level results were then combined to form note-level results.
   - Classification at the note level was defined as “yes” if at least 1 sentence in the note was marked “yes.” Otherwise, it was classified as “no”.

The diagnosis of DVT itself was not considered a DVT symptom. Additionally, the bodily location (ie, source) of pain was considered when determining the presence of any symptom (such as DVT, back pain, or abdominal or epigastric pain). For example, pain radiating from the upper or lower extremity was considered a DVT symptom, whereas pain radiating to the upper or lower extremity was not. Similarly, pain that radiated to the back region was not counted as back pain, and pain that radiated to the abdomen or epigastric region was not counted as abdominal or epigastric pain.

Performance Evaluation

The results of the NLP algorithm against the validation data set were compared to the adjudicated chart review results notes. For each symptom, the numbers of true positive (TP), false positive (FP), true negative (TN), and false negative (FN) cases were used to estimate the sensitivity or recall, specificity, PPV or precision, negative predictive value (NPV), and overall $F_1$-score, a harmonic balance measurement of PPV and
sensitivity. Sensitivity was defined as the number of TP s divided by the total number of symptoms ascertained by the chart reviews (TP+FN). PPV was defined as the number of TP s divided by the total number of symptoms identified by the computerized algorithm (TP+FP). Specificity was defined as the number of TN s divided by the total number of notes without symptoms ascertained by the chart reviews (TN+FP). NPV was defined as the number of TN s divided by the total number of notes identified by the computerized algorithm without symptoms (TN+FN). The $F_1$-score was calculated as $(2 \times \text{PPV} \times \text{sensitivity}) / (\text{PPV} + \text{sensitivity})$.

Interrater Reliability Among 2 Annotators
The agreement and kappa coefficient against the double-annotated subset were calculated to assess the interrater reliability among the annotators.

Discrepancy Analysis
For each symptom, discordant results between the NLP algorithm and adjudicated chart review against the validation data set were analyzed. Both FP and FN scenarios were summarized in detail.

Implementation of the NLP Algorithm
The validated computerized algorithm was implemented via Python programming on a Linux server to process the qualified study notes with the exception of training and validation notes. For each symptom, the process created the results of each note at the sentence level and note level for summary analysis.

Results

Statistics of the Study Notes
A total of 408,147 and 709,789 notes were retrieved for 2611 PDAC cases and 10,085 matched controls, respectively. The distribution of the notes and patient demographics are summarized in Table 1. Compared to patients without PDAC, patients with PDAC were older and more likely to be men (PDAC cases: mean 69.2, SD 9.1 years of age and n=1328, 50.9% men; controls: mean 48.6, SD 17.2 years of age and n=4681, 46.4% men). A total of 3,827,166 sentences and 69,455,767 word tokens were derived from notes belonging to patients with PDAC. The corresponding numbers were 5,880,717 sentences and 102,358,031 word token for patients without PDAC. Both the average number of notes per patient and average words per note were higher for patients with PDAC (notes per patient: mean 156.3, SD 138.3; words per note: mean 170.2, SD 319.2) compared to patients without PDAC (notes per patient: mean 70.4, SD 94.1; words per note: mean 144.2, SD 263.6).

Table 1. Description of the study population and the associated data sets.

<table>
<thead>
<tr>
<th></th>
<th>PDAC$^a$ (n=2611)</th>
<th>Non-PDAC (n=10,085)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Age (years), mean (SD)</td>
<td>69.2 (9.1)</td>
<td>48.6 (17.2)</td>
</tr>
<tr>
<td>Gender: women, n (%)</td>
<td>1283 (49.1)</td>
<td>5404 (53.6)</td>
</tr>
<tr>
<td>Gender: men, n (%)</td>
<td>1328 (50.9)</td>
<td>4681 (46.4)</td>
</tr>
<tr>
<td>Total clinical notes, n</td>
<td>408,147</td>
<td>709,789</td>
</tr>
<tr>
<td>Total sentences, n</td>
<td>3,827,166</td>
<td>5,880,717</td>
</tr>
<tr>
<td>Total word tokens, n</td>
<td>69,455,767</td>
<td>102,358,031</td>
</tr>
<tr>
<td>Notes per patient, mean (SD)</td>
<td>156.3 (138.3)</td>
<td>70.4 (94.1)</td>
</tr>
<tr>
<td>Sentences per clinical note, mean (SD)</td>
<td>9.4 (15.7)</td>
<td>8.3 (13.9)</td>
</tr>
<tr>
<td>Words per clinical note, mean (SD)</td>
<td>170.2 (319.2)</td>
<td>144.2 (263.6)</td>
</tr>
</tbody>
</table>

$^a$PDAC: pancreatic ductal adenocarcinoma.

Interrater Reliability of 2 Annotators
The agreement and kappa coefficient between 2 annotators for a subset of notes (n=2795) is summarized in Table S5 in Multimedia Appendix 1. The agreement ranged from 98.82% (abdominal or epigastric pain) to 99.96% (upper extremity DVT), while the kappa coefficient ranged from 0.6 (insomnia) to 0.91 (abdominal or epigastric pain).

Validation of the NLP Algorithm
Table 2 summarizes the performance of the computerized NLP algorithm against the adjudicated chart review results of 1000 notes based on the validation data set. In descending order, the precision (PPV) of the algorithms ranged from 98.9% (jaundice) to 84% (lower extremity DVT), recall (sensitivity) ranged from 98.1% (weight loss) to 82.8% (epigastric bloating), specificity ranged from 99.9% (epigastric bloating, jaundice, and pruritus) to 98.9% (depression), NPV ranged from 99.9% (lower extremity DVT) to 98.1% (abdominal or epigastric pain and back pain), and the $F_1$-score ranged from 0.97 (jaundice) to 0.87 (depression).
Table 2. The computerized model’s performance against the adjudicated chart review results in the validation data set (n=1000).

<table>
<thead>
<tr>
<th>Symptoms</th>
<th>TP (n)</th>
<th>TN (n)</th>
<th>FP (n)</th>
<th>FN (n)</th>
<th>Sensitivity (%)</th>
<th>PPV (%)</th>
<th>Specificity (%)</th>
<th>NPV (%)</th>
<th>F1-score</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Gastrointestinal symptoms</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Abdominal or epigastric pain</td>
<td>156</td>
<td>824</td>
<td>4</td>
<td>16</td>
<td>90.7</td>
<td>97.5</td>
<td>99.5</td>
<td>98.1</td>
<td>0.94</td>
</tr>
<tr>
<td>Anorexia or early satiety</td>
<td>78</td>
<td>909</td>
<td>2</td>
<td>11</td>
<td>87.6</td>
<td>97.5</td>
<td>99.8</td>
<td>98.8</td>
<td>0.92</td>
</tr>
<tr>
<td>Dark urine</td>
<td>51</td>
<td>938</td>
<td>3</td>
<td>8</td>
<td>86.4</td>
<td>94.4</td>
<td>99.7</td>
<td>99.2</td>
<td>0.90</td>
</tr>
<tr>
<td>Epigastric bloating</td>
<td>53</td>
<td>935</td>
<td>1</td>
<td>11</td>
<td>82.8</td>
<td>98.2</td>
<td>99.9</td>
<td>98.8</td>
<td>0.90</td>
</tr>
<tr>
<td>Nausea or vomiting</td>
<td>97</td>
<td>820</td>
<td>3</td>
<td>7</td>
<td>93.3</td>
<td>97</td>
<td>99.6</td>
<td>99.2</td>
<td>0.95</td>
</tr>
<tr>
<td>Pale stool</td>
<td>40</td>
<td>949</td>
<td>5</td>
<td>6</td>
<td>87</td>
<td>88.9</td>
<td>99.5</td>
<td>99.4</td>
<td>0.88</td>
</tr>
<tr>
<td><strong>Systemic symptoms</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Back pain</td>
<td>95</td>
<td>882</td>
<td>6</td>
<td>17</td>
<td>84.8</td>
<td>94.1</td>
<td>99.3</td>
<td>98.1</td>
<td>0.89</td>
</tr>
<tr>
<td>Fatigue</td>
<td>105</td>
<td>883</td>
<td>2</td>
<td>10</td>
<td>91.3</td>
<td>98.1</td>
<td>99.8</td>
<td>98.9</td>
<td>0.95</td>
</tr>
<tr>
<td>Jaundice</td>
<td>90</td>
<td>905</td>
<td>1</td>
<td>4</td>
<td>95.7</td>
<td>98.9</td>
<td>99.9</td>
<td>99.6</td>
<td>0.97</td>
</tr>
<tr>
<td>Malaise</td>
<td>52</td>
<td>941</td>
<td>2</td>
<td>5</td>
<td>91.2</td>
<td>96.3</td>
<td>99.8</td>
<td>99.5</td>
<td>0.94</td>
</tr>
<tr>
<td>Pruritus</td>
<td>27</td>
<td>970</td>
<td>1</td>
<td>2</td>
<td>93.1</td>
<td>96.4</td>
<td>99.9</td>
<td>99.8</td>
<td>0.95</td>
</tr>
<tr>
<td>Weight loss</td>
<td>101</td>
<td>886</td>
<td>11</td>
<td>2</td>
<td>98.1</td>
<td>90.2</td>
<td>99.8</td>
<td>99.8</td>
<td>0.94</td>
</tr>
<tr>
<td><strong>Mental symptoms</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Anxiety</td>
<td>79</td>
<td>911</td>
<td>3</td>
<td>7</td>
<td>91.9</td>
<td>96.3</td>
<td>99.7</td>
<td>99.2</td>
<td>0.94</td>
</tr>
<tr>
<td>Depression</td>
<td>83</td>
<td>892</td>
<td>10</td>
<td>15</td>
<td>84.7</td>
<td>89.3</td>
<td>98.9</td>
<td>98.3</td>
<td>0.87</td>
</tr>
<tr>
<td>Insomnia</td>
<td>62</td>
<td>925</td>
<td>7</td>
<td>6</td>
<td>91.2</td>
<td>89.9</td>
<td>99.3</td>
<td>99.4</td>
<td>0.91</td>
</tr>
<tr>
<td><strong>Vascular conditions</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Lower extremity DVT(^h) symptom</td>
<td>19</td>
<td>977</td>
<td>3</td>
<td>1</td>
<td>95</td>
<td>86.4</td>
<td>99.7</td>
<td>99.9</td>
<td>0.91</td>
</tr>
<tr>
<td>Upper extremity DVT symptom</td>
<td>21</td>
<td>972</td>
<td>4</td>
<td>3</td>
<td>87.5</td>
<td>84</td>
<td>99.6</td>
<td>99.7</td>
<td>0.86</td>
</tr>
</tbody>
</table>

\(^a\)TP: true positive. 
\(^b\)TN: true negative. 
\(^c\)FP: false positive. 
\(^d\)FN: false negative. 
\(^e\)PPV: positive predicted value. 
\(^f\)NPV: Negative predicted value. 
\(^g\)Hospital encounter notes were excluded with the exception of emergency notes. 
\(^h\)DVT: deep vein thrombosis.

**Discrepancy Analysis**

The discrepancy analysis is summarized in Table S6 in Multimedia Appendix 1. The most common scenarios that resulted in FPs were failure of exclusion of the symptoms described in the patient medical problem list, failure of exclusion of symptoms from instructions, failure of negation, or failure of exclusion of a symptom from past medical history. The most common scenarios for FNs were false negation, missing specific terms or patterns of terms in the search list, false classification of past history symptoms, or false exclusion of symptoms described in relevant medication instructions.

**Implementation of the NLP Algorithm**

Table 3 summarizes the symptoms identified by the validated NLP algorithms based on the implementation data set. Of the 393,003 and 708,489 notes belonging to PDAC and non-PDAC patients, respectively, at least 1 symptom was identified in 52,803 (13.44%) and 56,552 (7.98%) notes, respectively. The presence of symptoms ranged (in descending order) from 4.98% (abdominal or epigastric pain) to 0.05% (upper extremity DVT) in patients with PDAC and from 1.75% (back pain) to 0.01% (pale stool) in the patients without PDAC.
Table 3. Presence of symptoms identified by the computerized algorithms based on the implementation data set at the clinical note level.

<table>
<thead>
<tr>
<th>Symptom</th>
<th>Clinical notes from patients with PDAC(^a), n (%) (n=393,003)</th>
<th>Clinical notes from patients without PDAC, n (%) (n=708,489)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Any of 17 symptoms</strong></td>
<td>52,803 (13.44)</td>
<td>56,552 (7.98)</td>
</tr>
<tr>
<td><strong>Gastrointestinal symptoms</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Abdominal or epigastric pain</td>
<td>19,582 (4.98)</td>
<td>11,274 (1.59)</td>
</tr>
<tr>
<td>Anorexia or early satiety</td>
<td>4393 (1.12)</td>
<td>1626 (0.23)</td>
</tr>
<tr>
<td>Dark urine</td>
<td>1511 (0.38)</td>
<td>121 (0.02)</td>
</tr>
<tr>
<td>Epigastric bloating</td>
<td>3217 (0.82)</td>
<td>1665 (0.24)</td>
</tr>
<tr>
<td>Nausea or vomiting</td>
<td>7754 (1.97)</td>
<td>7429 (1.05)</td>
</tr>
<tr>
<td>Pale stool</td>
<td>875 (0.22)</td>
<td>35 (0.01)</td>
</tr>
<tr>
<td><strong>Systemic symptoms</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Back pain</td>
<td>8407 (2.14)</td>
<td>12,416 (1.75)</td>
</tr>
<tr>
<td>Fatigue</td>
<td>7170 (1.82)</td>
<td>9621 (1.36)</td>
</tr>
<tr>
<td>Jaundice</td>
<td>9118 (2.32)</td>
<td>305 (0.04)</td>
</tr>
<tr>
<td>Malaise</td>
<td>2984 (0.76)</td>
<td>4162 (0.59)</td>
</tr>
<tr>
<td>Pruritus</td>
<td>1872 (0.48)</td>
<td>622 (0.09)</td>
</tr>
<tr>
<td>Weight loss</td>
<td>8001 (2.04)</td>
<td>2619 (0.37)</td>
</tr>
<tr>
<td><strong>Mental symptoms</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Anxiety</td>
<td>3924 (1)</td>
<td>10,843 (1.53)</td>
</tr>
<tr>
<td>Depression</td>
<td>4995 (1.27)</td>
<td>10,810 (1.53)</td>
</tr>
<tr>
<td>Insomnia</td>
<td>2228 (0.57)</td>
<td>4159 (0.59)</td>
</tr>
<tr>
<td><strong>Vascular conditions</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Lower extremity DVT(^b) symptom</td>
<td>807 (0.21)</td>
<td>1465 (0.21)</td>
</tr>
<tr>
<td>Upper extremity DVT symptom</td>
<td>215 (0.05)</td>
<td>719 (0.1)</td>
</tr>
</tbody>
</table>

\(^a\)PDAC: pancreatic ductal adenocarcinoma.

\(^b\)DVT: deep vein thrombosis.

**Discussion**

In this study, we developed computerized NLP algorithms to identify 17 symptoms that were documented prior to PDAC diagnosis from clinical notes and patient-provider communication emails. To our knowledge, this is the first study to systematically identify a set of symptoms related to PDAC using NLP. When assessed against the manually annotated results, the algorithm achieved a reasonable performance, with recall (sensitivity) ranging from 82.6% to 98.1% and precision (PPV) ranging from 84% to 98.9%.

Accurate extraction of symptoms embedded in free-text notes posed a significant challenge. First, the symptoms might be described in various portions of the notes. For example, symptoms might be embedded under past medical history, review of systems, the patient’s medical problem list, instructions, sign and symptom warnings, questionnaires, checklists, lab orders and tests, medications, procedures, diagnosis, or chief complaints. Second, health care providers might copy and paste information from previous notes. In addition, we would like to highlight some specific challenges. First, a negated term could sometimes apply to only 1 symptom or to multiple symptoms after negation (eg, no coughing, no chest pain, no abdomen pain; denies nausea or vomiting, diarrhea, constipation, abdominal pain). Second, the defined rules might not address all scenarios. For example, one of our defined rules for abdominal pain required the word “pain” and the body location to be within a 5-word distance. If the words for body location (eg, abdomen) and “pain” were separated by more than 5 words, the sentence was marked “no” for abdominal pain. Third, we found that some symptom terms could have different meanings, which caused FPs. For example, the phrase “lower bp” for back pain could also mean lower blood pressure, and the fatigue term “exhausted” could refer to either physical or mental exhaustion. Fourth, some exclusion criteria, as shown in Table S3 in Multimedia Appendix 1 (eg, exclude localized itching for pruritus), also caused potential misclassification.

The data annotation process was tedious and time-consuming. The following lessons learned could benefit the medical research community. First, set up a training period for chart annotators and study investigators with medical backgrounds to review at least several hundred notes (the same notes for all the annotators). This step would not only allow the chart annotators to...
to be trained for the process but also would identify potential issues that might arise during the formal review process. Second, develop a chart annotation document that would include the detailed inclusion and exclusion criteria to be used for the annotation. The document should define specific types of notes (eg, mental health progress notes) or sections of the notes (eg, “past medical history” or “history of present illness”) to be reviewed or to be skipped. The document should also outline rules to determine the presence or absence of the conditions of interest. For example, if a patient experienced abdominal pain at home but did not experience pain at the time of the visit. Such rules are study-specific, but they need to be considered thoroughly and documented.

Advanced transformer language models, including bidirectional encoder representations from transformers (BERT) [20], clinical BERT [34], BioBERT [35], and BERT for EHRs (BEHRT) [36], have gained popularity in research involving NLP. These NLP language models offer the advantage of contextual understanding through embedding representations, allowing the developed algorithms to capture the meaning and intricate relationships within the text and enhance the accuracy of the analysis. They have been widely used for analyzing information from unstructured notes in the healthcare domain [18,19,37]. Research in this area in future work is warranted to further boost the performance of PDAC-related symptoms, especially for these lower performances via the rule-based approach.

Our study acknowledged several potential limitations. First, the completeness and accuracy of the extracted symptoms depended on the information documented in the EHR system. Incomplete or inaccurate documentation of symptoms could lead to bias. Second, although our training process was quite comprehensive and included a relatively large number of notes, the rules and lexicons built based on the training data sets were still not highly comprehensive, as summarized in the discrepancy analysis. Therefore, a more extensive sample could be used to enhance the rules and lexicons if applied in other populations in the future, especially for rare symptoms. Third, a few terms or phrases could indicate meanings other than the symptom of interest (eg, “patient has exhausted all conservative measures” or “patient complaint of lower bp than usual”). Additional contexts with these terms would be required to determine the actual meaning. Fourth, for symptoms involving body location, such as abdominal pain and back pain, the allowed distance between the location and the symptom could sometimes lead to the misclassification of TP cases. Lastly, when applied to other healthcare systems and settings, the developed computerized algorithms might require modifications due to variations in the format and presentation of clinical notes in different healthcare settings.

In conclusion, the developed computerized algorithm and process could effectively identify relevant symptoms prior to PDAC diagnosis based on unstructured notes in a real-world care setting. This algorithm and process could be used to support the early detection of pancreatic cancer if implemented within a healthcare system to automatically identify patients with PDAC-related symptoms, especially those with PDAC-specific symptoms.

Acknowledgments
This study was supported by The Pancreatic Cancer Action Network. The opinions expressed are solely those of the authors and do not necessarily reflect the official views of the funding agency. The authors thank the survey participants from the Consortium for the Study of Pancreatitis, Diabetes, and Pancreatic Cancer working group to determine the PDAC-related symptoms. The authors thank the patients of Kaiser Permanente Southern California for helping to improve care through the use of information collected through our electronic health record systems.

Conflicts of Interest
None declared.

Multimedia Appendix 1
Supplemental tables.
[DOCX File, 51 KB - ai_v3i1e51240_app1.docx]

References


Abbreviations

- BERT: bidirectional encoder representations from transformers
- DVT: deep vein thrombosis
- EHR: electronic health record
- FN: false negative
- FP: false positive
- KPSC: Kaiser Permanente Southern California
- NLP: natural language processing
- NPV: negative predictive value
- PDAC: pancreatic ductal adenocarcinoma
- PPV: positive predictive value
- TN: true negative
- TP: true positive

Edited by K El Emam, B Malin; submitted 26.07.23; peer-reviewed by B Senst, M Elbattah, Y Khan; comments to author 17.11.23; revised version received 08.12.23; accepted 16.12.23; published 15.01.24.

Please cite as:

Xie F, Chang J, Luong T, Wu B, Lustigova E, Shrader E, Chen W
Identifying Symptoms Prior to Pancreatic Ductal Adenocarcinoma Diagnosis in Real-World Care Settings: Natural Language Processing Approach
JMIR AI 2024;3:e51240
URL: https://ai.jmir.org/2024/1/e51240
doi:10.2196/51240
PMID:
Beyond the Hype—The Actual Role and Risks of AI in Today’s Medical Practice: Comparative-Approach Study

Steffan Hansen¹, MA; Carl Joakim Brandt¹, PhD; Jens Søndergaard¹, PhD
Research Unit of General Practice, Institution of Public Health, University of Southern Denmark, Odense, Denmark

Corresponding Author:
Steffan Hansen, MA
Research Unit of General Practice
Institution of Public Health
University of Southern Denmark
J.B. Winsløws Vej 9
Odense, 5000
Denmark
Phone: 45 65 50 36 19
Email: sholsthansen@health.sdu.dk

Abstract

Background: The evolution of artificial intelligence (AI) has significantly impacted various sectors, with health care witnessing some of its most groundbreaking contributions. Contemporary models, such as ChatGPT-4 and Microsoft Bing, have showcased capabilities beyond just generating text, aiding in complex tasks like literature searches and refining web-based queries.

Objective: This study explores a compelling query: can AI author an academic paper independently? Our assessment focuses on four core dimensions: relevance (to ensure that AI’s response directly addresses the prompt), accuracy (to ascertain that AI’s information is both factually correct and current), clarity (to examine AI’s ability to present coherent and logical ideas), and tone and style (to evaluate whether AI can align with the formality expected in academic writings). Additionally, we will consider the ethical implications and practicality of integrating AI into academic writing.

Methods: To assess the capabilities of ChatGPT-4 and Microsoft Bing in the context of academic paper assistance in general practice, we used a systematic approach. ChatGPT-4, an advanced AI language model by Open AI, excels in generating human-like text and adapting responses based on user interactions, though it has a knowledge cut-off in September 2021. Microsoft Bing’s AI chatbot facilitates user navigation on the Bing search engine, offering tailored search results.

Results: In terms of relevance, ChatGPT-4 delved deeply into AI’s health care role, citing academic sources and discussing diverse applications and concerns, while Microsoft Bing provided a concise, less detailed overview. In terms of accuracy, ChatGPT-4 correctly cited 72% (23/32) of its peer-reviewed articles but included some nonexistent references. Microsoft Bing’s accuracy stood at 46% (6/13), supplemented by relevant non-peer-reviewed articles. In terms of clarity, both models conveyed clear, coherent text. ChatGPT-4 was particularly adept at detailing technical concepts, while Microsoft Bing was more general. In terms of tone, both models maintained an academic tone, but ChatGPT-4 exhibited superior depth and breadth in content delivery.

Conclusions: Comparing ChatGPT-4 and Microsoft Bing for academic assistance revealed strengths and limitations. ChatGPT-4 excels in depth and relevance but falters in citation accuracy. Microsoft Bing is concise but lacks robust detail. Though both models have potential, neither can independently handle comprehensive academic tasks. As AI evolves, combining ChatGPT-4’s depth with Microsoft Bing’s up-to-date referencing could optimize academic support. Researchers should critically assess AI outputs to maintain academic credibility.

(JMIR AI 2024;3:e49082) doi:10.2196/49082

KEYWORDS
AI; artificial intelligence; ChatGPT-4; Microsoft Bing; general practice; ChatGPT; chatbot; chatbots; writing; academic; academia; Bing
Introduction

Artificial intelligence’s (AI) journey has been nothing short of incredible. Starting with its early days of rule-based systems, we have seen it grow and mature, stepping into the realm of machine learning, and more recently, diving into deep learning. This transformative journey has shaken up a lot of sectors, but health care is where AI has truly left an indelible mark.

Today, algorithms can spot issues in our x-rays or magnetic resonance imaging, sometimes even better than our seasoned doctors [1]. AI does not just stop there; it even gives us a heads-up on potential life-threatening situations in intensive care units, predicting conditions like septic shock hours before they occur. The world of drug discovery is moving faster than ever, thanks to AI’s helping hand [2]. However, as with most things, there are issues. There are big questions about how we protect our data and ensure different health record systems talk to each other [3], not to mention the lingering worries about biases in AI and the sometimes uneasy feeling of trusting a machine we do not fully “get” [4].

When you look at the big picture, we see ground-breaking models like GPT-3, ChatGPT-4 [5,6], and Microsoft Bing [7] making waves. They are not just about churning out text. They are doing things we had never imagined, like assisting in literature searches or refining our everyday web-based searches [8]. Their accomplishments in challenges, such as the Turing Test [9] and the LAMBADA (LAnguage Modeling Broadened to Account for Discourse Aspects) tasks [10], just go on to show how capable they are. Comparing powerhouses like ChatGPT-4 and Bing is not just for fun; it gives us a glimpse into where AI’s language abilities might be headed, and with new kids on the block like Google Bard, the sky is the limit [11]. Writing an academic paper, though? That is still a world where the human touch shines. From combing through mountains of literature to connecting the dots in innovative ways, it is a craft that demands the very best of us, but here is a thought: given how far AI has come, could it, one day, pen down an academic masterpiece on its own? This paper is all about that tantalizing question.

As we embark on this exploration, we will keenly assess a few critical dimensions:

- Relevance: can AI ensure that its response precisely addresses the prompt and brings to the table information that is truly pertinent to the question or topic?
- Accuracy: how reliable is AI in delivering information that is not just factually correct but also up-to-date with the current pulse of the academic field?
- Clarity: when we read what is written by AI, does it resonate with clarity, coherence, and a logical flow of ideas, all presented with precise and unambiguous language?
- Tone and style: given the seriousness of academic papers, can AI match the appropriate tone and style, ensuring it resonates with the formality and professionalism we expect to see in academic texts?

We are diving deep to see if AI can muster up the relevance, accuracy, clarity, and tone we associate with academic work, and of course, while we probe these questions, we are not losing sight of the overarching ethics and practicality of inviting AI into the revered domain of academic writing.

Methods

Ethical Considerations

In Denmark, ethical committee approval is only mandatory for studies that include trials involving liveborn human individuals, human gametes intended for fertilization, fertilized human eggs, embryonic cells and embryos, tissue, cells and genetic material from humans, embryos, etc, or deceased persons. Also included are clinical trials of medicines in humans and clinical trials of medical devices. Hence, our study did not require approval from an ethical committee.

Overview

In this methods section, we have detailed the approach taken to evaluate and compare the performance of ChatGPT-4 and Microsoft Bing in the context of assisting with an academic paper in the realm of general practice. This section outlines the data collection process, prompt design, evaluation criteria, and analysis of the AI-generated responses.

Models

ChatGPT-4

ChatGPT-4 is an advanced AI language model developed by OpenAI [5], based on the ChatGPT-4 architecture. It is designed to generate human-like text and engage in interactive conversations with users. Trained on a vast data set, ChatGPT-4 demonstrates a strong understanding of context, language, and reasoning abilities. When using GPT-4, it is important to highlight that during a conversation, the information and discussion are dynamically shaped throughout the interaction. Indeed, GPT-4 can respond by incorporating the information the user provides, potentially leading to different outcomes even for users with similar queries. This dynamic nature is crucial for understanding how a large language model like GPT-4 operates.

Although ChatGPT-4 can perform various tasks, such as answering questions, providing recommendations, and generating content, it has a knowledge cut-off date of September 2021. This means that the model has been trained on a data set consisting of text and information available up until that point. Therefore, any events, advancements, or changes in various fields that have occurred since September 2021 will not be known to ChatGPT-4. Additionally, it should be noted that ChatGPT-4, like any AI language model, reflects the data on which it has been trained. As a result, its knowledge might contain inaccuracies, biases, or outdated information even for events and topics within its known time frame.

Microsoft Bing

The Microsoft Bing AI chatbot [7] is an intelligent conversational agent developed by Microsoft Corporation, designed to assist users in navigating the Microsoft Bing search engine and answering various queries. Leveraging AI, natural language processing, and machine learning, the Microsoft Bing
AI chatbot understands user inputs and provides relevant information or search results accordingly. Integrated seamlessly with the Microsoft Bings platform, the chatbot offers a user-friendly and interactive way to engage with search functionalities, enhancing the overall user experience.

Prompt Design
In the context of AI, especially with large language models, a “prompt” refers to a set of instructions or a question given to the AI to guide its response. The purpose of a prompt is to set clear expectations for the AI’s output and to ensure that the response generated aligns with the user’s intent.

A prompt was designed to secure the AI models’ ability to understand and generate accurate, relevant, and coherent responses in a formal and professional tone. Each prompt provided the AI models with the context of an academic paper and set the tone and expectations for the responses. The following specific prompt was used to ensure that both ChatGPT-4 and Microsoft Bing were primed for the task at hand:

I need your help with an academic paper. Please provide me with clear and concise explanations, using evidence and logical reasoning to support your responses. Your tone should be formal and professional, and your language should be free from errors and ambiguity. I am looking for accurate and well-supported information that will help me to achieve my academic goals.

Data Collection
The interview with the 2 models took place on March 9, 2023, with early access to ChatGPT-4. Both ChatGPT-4 and Microsoft Bing were asked to provide an outline for a discussion article on the chosen topic, encompassing various aspects of general practice. This approach aimed to evaluate the AI models’ ability to synthesize information and structure a coherent, well-organized outline that could serve as a foundation for a comprehensive discussion article. As differences between the outlines are likely, the most comprehensive outline was used to ensure a meaningful comparison between interviews. The length of each question was limited to ensure accuracy and reduce the risk of errors during the conversation.

Evaluation Criteria
It is important to note that the evaluation was conducted solely by one author, and the assessments were largely based on their subjective judgment. To compare and assess the quality of the AI-generated responses, the following evaluation criteria were established:

- **Relevance**: the extent to which the AI-generated response addresses the prompt and provides information pertinent to the question or topic.
- **Accuracy**: the degree to which the information provided is factually correct and up to date, based on the current state of knowledge in the field.
- **Clarity**: the clarity and coherence of the AI-generated response, including the logical flow of ideas and the use of precise, unambiguous language.
- **Tone and style**: the appropriateness of the tone and style of the AI-generated response, considering the formal and professional context of an academic paper.

To evaluate the evaluation criteria, a comprehensive literature search was conducted to identify areas where AI might be useful and implemented in general practice.

Analysis
Each AI-generated response was analyzed independently, using the evaluation criteria, providing the strengths and weaknesses of each model. Hereafter, a comparison between the 2 models was conducted to establish differences. The results of the evaluation and comparison between the 2 models were then compiled and analyzed to determine the overall performance of ChatGPT-4 and Microsoft Bing related to the area of AI use in general practice and the areas preidentified, aiming at identifying the strengths and weaknesses of each AI model as well as any potential areas for improvement.

Results
For a complete comparison, the full conversation with both ChatGPT-4 and Microsoft Bing models can be found in Multimedia Appendix 1.

Relevance
**Chat-GPT**
GPT-4 offers a detailed analysis of AI applications in health care, focusing on general practice, its limitations, ethical concerns, and the importance of collaboration between AI and health care professionals. It provides comprehensive information, citing academic sources and studies, discussing AI algorithms, natural language processing, pattern recognition, evidence-based medicine, and personalized treatment plans. ChatGPT-4 also addresses data privacy, security concerns, and technical challenges while emphasizing the need to integrate AI systems with clinical workflows and patient needs. It provides a relevant and comprehensive examination of AI’s potential benefits and challenges in health care, emphasizing the need for integration with clinical workflows and a balanced approach to ensure optimal patient care.

**Microsoft Bing**
Microsoft Bing offers a brief overview of AI in general practice, addressing advantages and limitations without delving into specific applications or ethical considerations. It lacks the depth and citations and does not emphasize the importance of collaboration between AI and health care professionals. Although Microsoft Bing touches on themes that are relevant, it provides neither specific study references nor in-depth explanations, offering a more concise perspective (Table 1).
Table 1. Comparison of ChatGPT and Microsoft Bing in terms of topic relevance.

<table>
<thead>
<tr>
<th>Evaluation criteria</th>
<th>ChatGPT-4</th>
<th>Microsoft Bing</th>
</tr>
</thead>
</table>
| Relevance           | • A detailed analysis of AI applications in healthcare  
                     • Comprehensive information and citing academic sources  
                     • Emphasizing the need for integration with clinical workflows and a balanced approach to ensure optimal patient care |
|                     | A brief overview of AI in general practice  
                     • Lack of in-depth or specific study citations  
                     • Offering a more concise perspective |

*a*AI: artificial intelligence.

**Accuracy**

**ChatGPT**

ChatGPT-4 included 23 of 32 (72%) precise peer-reviewed articles with high accuracy. The introduction and applications in general practice were 100% correct. However, it also cited 9 nonexistent articles, with 4 out of 7 inaccuracies in limitations and all 4 ethical considerations being inaccurately cited.

**Microsoft Bing**

Microsoft Bing included 6 of 13 (46%) highly accurate, peer-reviewed articles, along with 7 non-peer-reviewed but highly relevant articles. Ethical considerations and applications in general practice cited 3 and 2 non-peer-reviewed articles, respectively (Table 2).

The references provided from both models, along with the accuracy distribution, can be found in Multimedia Appendix 2.

Table 2. Comparison of ChatGPT and Microsoft Bing in terms of accuracy.

<table>
<thead>
<tr>
<th>Evaluation criteria</th>
<th>ChatGPT-4</th>
<th>Microsoft Bing</th>
</tr>
</thead>
</table>
| Accuracy            | • A total of 23 out of 32 (72%) precise peer-reviewed articles, with high accuracy  
                     • A total of 9 nonexistent articles, with specific inaccuracies |
|                     | A total of 6 out of 13 (46%) highly accurate, peer-reviewed articles  
                     • A total of 7 non-peer-reviewed but highly relevant articles |

**Clarity**

**Chat GPT-4**

Overall, the text generated by ChatGPT demonstrates a high level of clarity and coherence, exhibiting a logical flow of ideas and the use of precise, unambiguous language. The text is easy to follow and understand, even for readers who may not be familiar with the technical terms and concepts discussed.

**Microsoft Bing**

Similar to ChatGPT, the text exhibits a high level of clarity and coherence, with a logical flow of ideas and the use of precise, unambiguous language. It is easily comprehensible, even for readers unfamiliar with the technical terms and concepts discussed. However, the text could be improved by providing more details and examples to support the points made, as many areas are discussed in a more general manner (Table 3).

Table 3. Comparison of ChatGPT and Microsoft Bing in terms of clarity.

<table>
<thead>
<tr>
<th>Evaluation Criteria</th>
<th>ChatGPT-4</th>
<th>Microsoft Bing</th>
</tr>
</thead>
<tbody>
<tr>
<td>Clarity</td>
<td>The text is clear, coherent, and easy to understand, even for nontechnical readers.</td>
<td></td>
</tr>
<tr>
<td></td>
<td>The text is clear and coherent but could benefit from more detailed examples.</td>
<td></td>
</tr>
</tbody>
</table>

**Tone (Chat GPT-4 and Microsoft Bing)**

Overall, the tone and style of the text are appropriate for the formal and professional context of an academic paper, effectively conveying complex ideas in a clear and objective manner (Table 4).

Table 4. Comparison of ChatGPT and Microsoft Bing in terms of tone.

<table>
<thead>
<tr>
<th>Evaluation criteria</th>
<th>ChatGPT-4</th>
<th>Microsoft Bing</th>
</tr>
</thead>
<tbody>
<tr>
<td>Tone</td>
<td>Appropriate for an academic paper, conveying ideas clearly and objectively</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Appropriate for an academic paper, conveying ideas clearly and objectively</td>
<td></td>
</tr>
</tbody>
</table>

**Discussion**

**Principal Findings**

In recent years, AI has become an increasingly prevalent tool in various domains, including health care and academic research. AI language models, such as ChatGPT-4 and Microsoft Bing, have demonstrated the potential to assist researchers in generating and organizing content for academic papers. In the context of general practice, a rapidly evolving field with a growing need for accurate and relevant information, understanding the strengths and limitations of these AI models is crucial for researchers and practitioners alike. This paper aimed to compare and analyze the performance of ChatGPT-4 and Microsoft Bing in assisting with an academic paper in general practice, focusing on their relevance, accuracy, clarity,
as well as tone and style. By examining their respective contributions and limitations, we seek to provide insights into their potential uses and areas for improvement in AI-assisted research.

In terms of relevance, ChatGPT-4 provided a detailed analysis of AI applications in health care, emphasizing the importance of collaboration between AI and health care professionals, while Microsoft Bing offered a concise overview without delving into specific applications or ethical considerations. As for accuracy, ChatGPT-4 accurately cited 72% (23/32) of peer-reviewed articles, but it also inaccurately cited 9 nonexistent articles. Microsoft Bing, on the other hand, included 6 of 13 (46%) accurate peer-reviewed articles and 7 non–peer-reviewed but highly relevant articles.

Regarding clarity, both ChatGPT-4 and Microsoft Bing demonstrated high levels of clarity and coherence, presenting a logical flow of ideas with precise, unambiguous language. Nevertheless, Microsoft Bing could benefit from providing more details and examples to support its points, as certain areas were discussed in a more general manner. Lastly, in terms of tone and style, both AI models used an appropriate tone and style for the formal and professional context of an academic paper, effectively conveying complex ideas in a clear and objective manner.

Comparison With the Existing Literature

The results of this study, which compared the performance of ChatGPT-4 and Microsoft Bing in assisting with an academic paper in general practice, can be contextualized within the broader landscape of AI applications in health care and general practice research. The findings align with several previous studies that have highlighted the potential of AI language models, such as ChatGPT-4, to deliver relevant, detailed, and coherent information on complex subjects like health care [6,12].

The superior performance of ChatGPT-4 in providing comprehensive and in-depth analysis aligns with its advanced architecture and extensive training on a vast data set, which has been documented to enable the model to generate human-like text and engage in interactive conversations with users [12]. Similarly, the results are consistent with previous research that has emphasized the importance of collaboration between AI and health care professionals to achieve optimal patient care [13].

However, the observed weaknesses in ChatGPT-4’s accuracy, specifically in citing nonexistent articles, highlight the limitations of AI language models in some areas of academic research. This issue has been acknowledged in existing literature, where concerns have been raised about the potential for AI-generated content to include inaccuracies, biases, or misinformation [14].

In contrast, Microsoft Bing’s more concise approach to providing information echoes its primary function as a search engine assistant rather than a specialized AI language model. This result is consistent with the notion that AI chatbots, while capable of providing relevant information, may not always deliver the depth and detail required for more demanding academic tasks [15].

Strengths

This study has some strengths, as follows:

- Prompt design: the study used a well-crafted prompt to ensure that both ChatGPT-4 and Microsoft Bing were primed for the task, which helped in generating accurate, relevant, and coherent responses in a formal and professional tone.
- Evaluation criteria: the established evaluation criteria (relevance, accuracy, clarity, as well as tone and style) provided a comprehensive framework for comparing and assessing the quality of the AI-generated responses.
- Analysis: the independent analysis of each AI-generated response, followed by a comparison between the 2 models, allowed for a thorough understanding of the strengths and weaknesses of each AI model.

Weaknesses

The weaknesses of the study are the following:

- Data collection: the study’s data collection method, which involved interviewing the 2 models, may have been limited in scope. A more comprehensive approach involving a larger sample of questions or topics could have provided a broader understanding of the AI models’ capabilities.
- Knowledge cut-off: ChatGPT-4 has a knowledge cut-off date of September 2021, which may have limited its ability to provide up-to-date information in some instances.
- Limited exploration of AI models: the study only compared 2 AI models—ChatGPT-4 and Microsoft Bing. This may not provide a complete picture of the landscape of AI tools available for assisting with academic papers in general practice. Including more AI models, such as Google’s chatbot—Bard, in the comparison could have yielded a more comprehensive analysis. However, this model is not currently available in Denmark.

The strengths and weaknesses of each model are presented in Table 5.
Table 5. A side-by-side comparison of the features and aspects of ChatGPT-4 and Microsoft Bing’s artificial intelligence (AI) chatbot.

<table>
<thead>
<tr>
<th>Feature or aspect</th>
<th>ChatGPT-4</th>
<th>Microsoft Bing</th>
</tr>
</thead>
<tbody>
<tr>
<td>Developer</td>
<td>OpenAI</td>
<td>Microsoft Corporation</td>
</tr>
<tr>
<td>Primary function</td>
<td>Generating human-like text and engaging in interactive conversations</td>
<td>Assisting users in navigating the Microsoft Bing search engine and answering queries</td>
</tr>
<tr>
<td>Training or technology</td>
<td>Vast data set, context understanding, language, and reasoning abilities</td>
<td>Artificial intelligence, natural language processing, and machine learning</td>
</tr>
<tr>
<td>Special features</td>
<td>Answering questions, providing recommendations, and generating content</td>
<td>Integrating with the Bing platform and enhancing the search experience</td>
</tr>
<tr>
<td>Conversation limits</td>
<td>25 conversations per 3 hours</td>
<td>Limited to 20 prompts</td>
</tr>
<tr>
<td>Internet access</td>
<td>No</td>
<td>Yes</td>
</tr>
<tr>
<td>Knowledge cut-off</td>
<td>Up to 2021</td>
<td>Uses OpenAI technology with access to the internet and thus can acquire the newest information</td>
</tr>
<tr>
<td>Memory constraints</td>
<td>Forgets information within longer conversations and might stop midsentence in lengthy responses</td>
<td>Closely related to ChatGPT-4 in this area</td>
</tr>
<tr>
<td>Additional information</td>
<td>Some responses may require user prompts to be complete</td>
<td>Offers a user-friendly and interactive way to engage with search functionalities</td>
</tr>
</tbody>
</table>

**Implications for AI-Assisted Research**

The findings of this study have several implications for researchers and practitioners using AI in general practice and other academic fields. These implications are as follows:

- **Quality of AI-generated content:** the comparison between ChatGPT-4 and Microsoft Bing demonstrates that the quality of AI-generated content can vary between models. Researchers and practitioners should be aware of the strengths and weaknesses of different AI models when selecting a tool to assist with their work.

- **Importance of collaboration:** both ChatGPT-4 and Microsoft Bing highlight the importance of collaboration between AI and health care professionals. AI systems should be designed to complement human expertise and foster collaboration, enhancing the overall quality of research and practice.

- **Relevance and accuracy:** ensuring the relevance and accuracy of AI-generated responses is crucial for researchers and practitioners. Although AI models can provide valuable insights, they might also generate inaccuracies or outdated information. Users must verify the information provided by AI models and cross-check it with up-to-date, reliable sources.

- **Clarity and tone:** AI-generated content should be clear and coherent; it should maintain an appropriate tone and style for the intended audience. Although AI models like ChatGPT-4 and Microsoft Bing show promising results in these aspects, users should carefully review and edit the generated content to ensure it meets the required standards.

- **Ethical considerations:** as AI continues to be integrated into various aspects of research and practice, ethical considerations must be addressed. Data privacy, security, and responsible use of AI-generated content are crucial to ensuring that AI is used responsibly and effectively in general practice and other academic fields.

Overall, the findings of this study indicate that AI models, such as ChatGPT-4 and Microsoft Bing, can provide valuable assistance in general practice and other academic fields. However, researchers and practitioners should be aware of the limitations and potential pitfalls of AI-generated content and use these tools thoughtfully and responsibly.

**Areas for Improvement and Future Research**

**AI Model Improvements**

**ChatGPT-4**

Although ChatGPT-4 demonstrates strong performance in relevance, clarity, and tone, there is room for improvement in terms of accuracy, especially in relation to citing nonexistent articles. Enhancing the fact-checking and source validation capabilities of the model could help address this issue.

**Microsoft Bing**

Microsoft Bing could benefit from improvements in providing more in-depth, relevant content with proper citations. Enhancing the model’s understanding of specific academic contexts and ethical considerations would allow it to provide more comprehensive and valuable insights to the users.

**Methodology Improvements**

The methodology improvements required are as follows:

- **Expanding the sample size:** including more AI models in the comparison would provide a broader understanding of the capabilities and limitations of AI-assisted research.

- **Diversifying the topics:** evaluating AI-generated responses across a wider range of topics and academic fields could offer more generalizable insights into the strengths and weaknesses of AI-assisted research.

- **Including human evaluation:** adding a panel of human evaluators to assess the AI-generated content could help provide a more nuanced understanding of the quality and relevance of the responses.
**Future Research Directions**

Some directions for future research are explained below:

- **Longitudinal studies:** investigating the evolution of AI models over time, as they are updated and trained on new data, could provide valuable insights into the progress of AI-assisted research and the potential of these tools in various academic fields.

- **Ethical implications:** examining the ethical implications of AI-generated content in academic research, such as issues related to plagiarism, data privacy, and potential biases, could help develop best practices and guidelines for responsible use of AI in research.

- **Integration with research workflows:** exploring how AI models can be effectively integrated into existing research workflows and practices and identifying the most effective ways to combine AI-generated content with human expertise would help maximize the benefits of AI-assisted research.

By addressing these areas for improvement and exploring future research directions, researchers and practitioners can continue to refine the use of AI models in general practice and other academic fields, ultimately enhancing the quality, efficiency, and impact of their work.

**Conclusions**

Our study comparing ChatGPT-4 and Microsoft Bing in assisting with writing an academic paper in general practice yielded several key findings. ChatGPT-4 demonstrated strong performance in terms of relevance, clarity, and tone, providing comprehensive information and detailed analysis of AI applications in health care. However, it exhibited weaknesses in accuracy, particularly in citing nonexistent articles. Microsoft Bing offered a more concise perspective, touching on relevant themes but lacking depth and proper citations.

In terms of methods used, the study incorporated prompt design, data collection, evaluation criteria, and analysis of AI-generated responses. The strengths of these methods include the design of a prompt that effectively engaged both AI models and the establishment of clear evaluation criteria. However, there is room for improvement in the methodology, such as expanding the sample size, diversifying the topics, and including human evaluation.

When comparing ChatGPT-4 and Microsoft Bing, ChatGPT-4 emerged as a more capable AI model for assisting with an academic paper in general practice. It provided a more in-depth, relevant, and coherent analysis of the topic; however, improvements in accuracy, particularly in source validation, would further enhance its utility. On the other hand, Microsoft Bing could benefit from improvements in providing more comprehensive content and proper citations to better support academic research.

In conclusion, ChatGPT-4 and Microsoft Bing present distinct pros and cons in academic writing. ChatGPT-4 excels in relevance and depth, but both AI models require improvement. Merging their strengths can produce comprehensive answers from ChatGPT-4 and up-to-date references from Microsoft Bing.

Despite their impressive abilities, these tools currently cannot author articles independently in certain areas. As AI models advance and incorporate current references and critical thinking, they may eventually conduct and create research autonomously. This study’s findings hold substantial implications for AI-assisted research across diverse fields, emphasizing areas for refinement and future research directions to optimize AI models in academia. To mitigate risks, researchers must adopt a critical approach, corroborate information from various sources, and stay aware of AI models’ limitations. This approach allows them to harness AI while preserving the integrity and rigor of their work.

**Conflicts of Interest**

None declared.

Multimedia Appendix 1
Interview with models.
[DOCX File, 50 KB - ai_v3i1e49082_app1.docx ]

Multimedia Appendix 2
References provided by models and their relevance.
[DOCX File, 22 KB - ai_v3i1e49082_app2.docx ]

**References**


5. GPT-4 is OpenAI’s most advanced system, producing safer and more useful responses. OpenAI. URL: https://openai.com/product/gpt-4 [accessed 2023-11-01]


Abbreviations

AI: artificial intelligence

LAMBADA: LAnguage Modeling Broadened to Account for Discourse Aspects

Edited by H Liu; submitted 17.05.23; peer-reviewed by M Salvagno, G Sebastian; comments to author 07.09.23; revised version received 11.10.23; accepted 15.10.23; published 22.01.24.

Please cite as:

Hansen S, Brandt CJ, Søndergaard J

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

JMIR AI 2024;3:e49082

URL: https://ai.jmir.org/2024/1/e49082
doi:10.2196/49082
PMID:

©Steffan Hansen, Carl Joakim Brandt, Jens Søndergaard. Originally published in JMIR AI (https://ai.jmir.org), 22.01.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR AI, is properly cited. The complete bibliographic information, a link to the original publication on https://www.ai.jmir.org/, as well as this copyright and license information must be included.

Jiahui Lu1,2, PhD; Huibin Zhang2, BE; Yi Xiao2, PhD; Yingyu Wang2, BE

1State Key Laboratory of Communication Content Cognition, People’s Daily Online, Beijing, China
2School of New Media and Communication, Tianjin University, Tianjin, China

Corresponding Author:
Huibin Zhang, BE
School of New Media and Communication
Tianjin University
Number 92, Weijin Road
Tianjin, 300072
China
Phone: 86 15135154977
Email: zhanghb@tju.edu.cn

Abstract

Background: Amidst the COVID-19 pandemic, misinformation on social media has posed significant threats to public health. Detecting and predicting the spread of misinformation are crucial for mitigating its adverse effects. However, prevailing frameworks for these tasks have predominantly focused on post-level signals of misinformation, neglecting features of the broader information environment where misinformation originates and proliferates.

Objective: This study aims to create a novel framework that integrates the uncertainty of the information environment into misinformation features, with the goal of enhancing the model’s accuracy in tasks such as misinformation detection and predicting the scale of dissemination. The objective is to provide better support for online governance efforts during health crises.

Methods: In this study, we embraced uncertainty features within the information environment and introduced a novel Environmental Uncertainty Perception (EUP) framework for the detection of misinformation and the prediction of its spread on social media. The framework encompasses uncertainty at 4 scales of the information environment: physical environment, macro-media environment, micro-communicative environment, and message framing. We assessed the effectiveness of the EUP using real-world COVID-19 misinformation data sets.

Results: The experimental results demonstrated that the EUP alone achieved notably good performance, with detection accuracy at 0.753 and prediction accuracy at 0.71. These results were comparable to state-of-the-art baseline models such as bidirectional long short-term memory (BiLSTM; detection accuracy 0.733 and prediction accuracy 0.707) and bidirectional encoder representations from transformers (BERT; detection accuracy 0.755 and prediction accuracy 0.728). Additionally, when the baseline models collaborated with the EUP, they exhibited improved accuracy by an average of 1.98% for the misinformation detection and 2.4% for spread-prediction tasks. On unbalanced data sets, the EUP yielded relative improvements of 21.5% and 5.7% in macro-F1-score and area under the curve, respectively.

Conclusions: This study makes a significant contribution to the literature by recognizing uncertainty features within information environments as a crucial factor for improving misinformation detection and spread-prediction algorithms during the pandemic. The research elaborates on the complexities of uncertain information environments for misinformation across 4 distinct scales, including the physical environment, macro-media environment, micro-communicative environment, and message framing. The findings underscore the effectiveness of incorporating uncertainty into misinformation detection and spread prediction, providing an interdisciplinary and easily implementable framework for the field.

(JMIR AI 2024;3:e47240) doi:10.2196/47240

KEYWORDS
misinformation detection; misinformation spread prediction; uncertainty; COVID-19; information environment
Introduction

Background
The World Health Organization and the United Nations have issued warnings about an “infodemic,” highlighting the spread of misinformation alongside the COVID-19 pandemic on social media [1]. Misinformation is characterized as “factually incorrect information not backed up by evidence” [2]. This misleading information frequently encompasses harmful health advice, misinterpretations of government control measures and emerging sciences, and conspiracy theories [3]. This phenomenon has inflicted detrimental impacts on public health, carrying “severe consequences with regard to people’s quality of life and even their risk of mortality” [4].

Automatic algorithms are increasingly recognized as valuable tools in mitigating the harm caused by misinformation. These techniques can rapidly identify misinformation, predict its spread, and have demonstrated commendable performance. The state-of-the-art detection techniques exhibit accuracy ranging from 65% to 90% [5,6], while spread-prediction techniques achieve performance levels between 62.5% and 77.21% [7,8].

The high accuracy of these techniques can be largely attributed to the incorporation of handcrafted or deep-learned linguistic and social features associated with misinformation [9-11]. Scholars have consistently invested efforts in integrating theoretically relevant features into algorithmic frameworks to enhance accuracy further.

Scholars have introduced diverse frameworks for misinformation detection and spread-prediction algorithms. Nevertheless, existing frameworks have predominantly concentrated on the intricate post-level signals of misinformation, emphasizing linguistic and social features (such as user relationships, replies, and knowledge sources) associated with misinformation. Notably, these frameworks have often overlooked the characteristics of the information environment in which misinformation originates and proliferates [12]. This neglect could potentially result in diminished performance for misinformation detectors when applied in various real-world misinformation contexts. This is due to the fact that different misinformation contexts possess unique characteristics within their information environment, influencing the types of misinformation that can emerge and thrive [13]. An indispensable characteristic of the information environment concerning misinformation is uncertainty. Uncertainty arises when the details of situations are ambiguous, complex, unpredictable, or probabilistic, and when information is either unavailable or inconsistent [14]. In uncertain situations, individuals tend to generate and disseminate misinformation as a means of resisting uncertainty and seeking understanding amid chaotic circumstances [15,16].

The COVID-19 pandemic serves as a notable example, marked by a lack of understanding of emerging science [17], uncertainties surrounding official guidelines and news reports [18], and unknown impacts on individuals and society [19]. Hence, in this study, we recognize uncertainty as the pivotal feature in the information environment of misinformation. Our objective is to formulate a novel framework for perceiving environmental uncertainty, specifically tailored for the detection and spread prediction of misinformation during the COVID-19 pandemic.

Our contributions can be outlined as follows. Theoretically, we provide a comprehensive exploration of uncertainty across 4 distinct scales of the information environment, namely, the physical environment, macro-media environment, micro-communicative environment, and message framing. These scales collectively contribute to the emergence and dissemination of misinformation. Furthermore, we hold the distinction of being the pioneers in integrating Environmental Uncertainty Perception (EUP) into the realms of misinformation detection and spread prediction. In terms of methodology, we introduce the EUP framework, designed to capture uncertainty signals from the information environment of a given post for both misinformation detection and spread prediction. Our experiments conducted on real-life data underscore the effectiveness of the EUP framework.

This paper unfolds as follows: In the “Related Work” section, we provide a concise review of the related work. The “Proposed Theoretical Framework” section elucidates uncertainty features within the information environment, which are pertinent to misinformation detection and spread prediction. Moving on to the “Research Objectives” section, we outline our study objectives. The “Methods” section details our methodology for testing the proposed framework. In the “Data Set and Experiment” section, we present our data set, experiments, and comprehensive analyses. The “Discussion” section delves into discussions on our findings, unraveling the theoretical and practical implications of our work. Finally, the “Conclusions” section concludes with a summary and outlines directions for future research.

Related Work
Detecting misinformation on social media represents a burgeoning research field that has garnered considerable academic attention. Multiple frameworks have been put forth for this task, primarily falling into 2 approaches: the post-only approach and the “zoom-in” approach [12]. In the former, frameworks focus on studying post features to differentiate misinformation from general information. Linguistic features, including novelty, complexity, emotions, and content topics, are frequently explored [6,11]. Additionally, researchers have delved into multimodal features, particularly those based on visuals [20,21]. Deep learning models in natural language processing have also proven beneficial for the misinformation detection task [5,22].

The “zoom-in” approach places emphasis on socio-contextual signals, centering on users’ networking aspects (eg, user relationships, number of replies, number of created threads; [23,24]) and network characteristics (eg, degree centrality [25]). Another line of research underscores the significance of relevant knowledge sources, including fact-checking websites [26] and knowledge graphs [27], which can be used to validate specific claims of interest.

Recently, Sheng et al [12] introduced a “zoom-out” approach, concentrating on the information environments of misinformation that can offer signals for detection. In their
approach, they incorporated the news environment into fake news detection. Their hypothesis posited that fake news should not only be relevant but also novel and distinct from recent popular news, enabling them to capture audience attention and achieve widespread dissemination. Their findings revealed that signals of popularity and novelty can enhance the performance of state-of-the-art misinformation detectors.

In the realm of misinformation detection, misinformation spread prediction represents another challenging task, albeit one that has received limited attention. This task involves predicting whether a piece of misinformation is likely to be disseminated to a broader audience through actions such as likes, comments, and shares. Within this context, our specific focus is on predicting whether misinformation is likely to be retweeted. This can be viewed as a binary classification task, akin to misinformation detection. Frameworks for this task typically incorporate linguistic and social features, which may overlap with or differ from those used in misinformation detection. Linguistic features such as persuasive styles, emotional expressions, and message coherence prove valuable in predicting the spread of misinformation [28,29]. Additionally, social features, including user metadata (e.g., number of friends, verification) and tweet metadata (e.g., presence of images and URLs), are identified as relevant factors for predicting misinformation spread [25].

Proposed Theoretical Framework

Uncertainty as a Central Aspect in Misinformation

Our study builds upon Sheng et al’s [12] “zoom-out” approach, adopting an interdisciplinary perspective that centers on the uncertainty within the information environment of misinformation. The realms of communication and psychology literature have conceptualized uncertainty as a fundamental aspect of misinformation. Uncertainty is said to prevail “when details of situations are ambiguous, complex, unpredictable, or probabilistic; uncertainty is also present when information is unavailable or inconsistent, and when individuals feel insecure about their own state of knowledge or the general state of knowledge” [14]. Confronted with uncertainty, individuals are driven to alleviate it by constructing their understanding of the situation [16]. This constructive process is known as sensemaking, which encompasses how individuals impart meaning to their surroundings and use it as a foundation for subsequent interpretation and action [30]. Sensemaking entails the utilization of information by individuals to fill gaps in their understanding [31]. Yet, the utilization of information in this manner does not always guarantee truth. In situations where information is slow to emerge, individuals are driven to comprehend uncertain situations by relying on their existing knowledge and heuristics for judgment. Unfortunately, this process often leads to the formation of false beliefs and misinformation [32]. Additionally, individuals may “turn to unofficial sources to satisfy their information needs,” potentially exposing themselves to inaccurate information [33]. As suggested by Kim et al [34], exposure to misinformation has the potential to diminish feelings of uncertainty. Moreover, as individuals integrate more information into their comprehension of a situation, there is a tendency to seek plausibility, which may lead to the generation and acceptance of misinformation [16,35].

The aforementioned tendencies are notably prominent in the context of the COVID-19 pandemic, as the pandemic represents a time of heightened uncertainty. The emergence of the pandemic was marked by a mysterious disease with previously unseen symptoms. Fundamental questions regarding the origins of the disease, measures for self-protection, and strategies for containing the outbreak were not immediately evident. As the pandemic progressed, uncertainty persisted regarding how and when the outbreak would be fully contained, as well as the long-term impact it would have on individuals and society. The uncertainty stemming from the pandemic, coupled with the surge of social media as a primary source of information, has facilitated the spread of misinformation [16].

Although many studies have identified “uncertainty” as a central aspect of misinformation, they have not thoroughly elucidated how uncertainty, as a crucial feature of the information environment, can aid in the detection of misinformation and the prediction of its spread. The literature frequently treats uncertainty as a static and holistic feature of a situation. However, the level of uncertainty within a situation can be dynamic, evolving as the situation progresses. For instance, uncertainties about the virus and the initial life changes induced by the COVID-19 pandemic would have been considerably higher at its onset than they are at present [36]. Moreover, uncertainty can manifest differently across various scales of the information environment. The information environment has become increasingly intricate with the proliferation of the internet and communication technologies. Individuals may be exposed to a substantial volume of information about trending topics through mainstream mass media (e.g., newspapers, TV, social media trends) within a short time frame, constituting a macro-media environment. Simultaneously, they may selectively engage in detailed communications on a specific issue provided by self-media (e.g., subscription accounts, self-broadcasting), shaping a micro-communicative environment. Uncertainty manifested in these 2 environments may independently or interactively influence people’s sensemaking processes and, consequently, their outputs (e.g., misinformation). Additionally, uncertainty can be inherent in the misinformation itself, providing cues for its detection and spread prediction. We will elaborate on the features of uncertainty in the information environment in the following section.

Uncertainty in the Information Environment

Uncertainty in the Physical Environment

Uncertainty prevails in the physical environment when unknown risks pose potential threats to our societal systems [15,16]. Scholars refer to such threats as “crises,” which can encompass natural disasters, large-scale accidents, social security incidents, and public health emergencies such as the pandemic [37]. Crises are marked by the existence of uncertainty and the imperative for timely decision-making [38]. Therefore, a crucial process during crises is sensemaking. However, the efforts needed for sensemaking will vary as a crisis progresses through stages. The Crisis and Emergency Risk Communication Model delineates 5 common stages in the crisis life cycle, spanning
“from risk, to eruption, to clean-up and recovery, and on into evaluation [38].” The eruption of the crisis, also known as the breakout stage, occurs when a key event triggers the crisis [39]. This is the period when the public becomes initially aware of the crisis, characterized by mysteries and heightened motivation to make sense of it. Evidence indicates that the breakout stage of a crisis harbors the highest level of uncertainty and demands extensive sensemaking efforts (eg, government updates [40]; social media communication [41]), consequently leading to a higher incidence of misinformation [42]. This evidence implies that misinformation is more likely to surface and proliferate in tandem with uncertainty in the information environment during the breakout stage compared with other stages throughout a crisis. These insights offer valuable cues for the detection and prediction of misinformation during the COVID-19 pandemic.

Uncertainty in the Macro-Media Environment

The macro-media environment encompasses recent media opinions and public attention to trending topics [12]. Governments and mainstream media play a pivotal role in setting the agenda for public attention. During crises such as the COVID-19 pandemic, governments frequently make swift and crucial decisions to safeguard the public. However, these decisions are often made without sufficient transparency, leading to potential uncertainties surrounding their rationale [43]. Such decisions inevitably draw media and public attention, quickly becoming trending topics in mainstream media outlets [44,45]. Regrettably, these rapid decisions often leave audiences with a high level of uncertainty about the reasons behind and the processes involved in making these decisions, potentially paving the way for misinformation. Supporting this notion, Lu [3] identified a correlation between the swift decision to quarantine Wuhan city and the emergence of misinformation regarding government control measures during the early stages of the COVID-19 pandemic in China. The evidence presented indicates that when public attention is directed toward a trending topic that carries uncertainty, misinformation is likely to emerge and spread. In simpler terms, it can be anticipated that when a piece of information is associated with a trending topic characterized by high uncertainty (as opposed to low uncertainty), there is a higher probability that the information could be misinformation and disseminated.

Uncertainty in the Micro-Communicative Environment

Differing from the macro-media environment, which offers a macro perspective on what mass audiences have recently read and focused on, the micro-communicative environment provides a micro view of the communication surrounding a specific issue. Both media and individuals tend to communicate using frames or terms imbued with uncertainty when discussing matters that lack evidence or consensus, such as those stemming from emerging science during the COVID-19 pandemic [32,46]. As an illustration, in the initial phase of the pandemic, when Hong Kong officials reported the first instance of a dog testing “weakly positive” for COVID-19 infection, subsequent media reports highlighted that “Hong Kong scientists aren’t sure [emphasis added] if the dog is actually infected or if it picked up the virus from a contaminated surface [47].” Experimental evidence has shown that such uncertainty frames about scientific matters can diminish people’s trust in science [48]. Empirical evidence from real-life social media data further indicates that a communication style marked by ambiguity can potentially lead audiences to generate and disseminate misinformation [32]. This body of findings implies that if information is embedded in uncertain (as opposed to consensus) communication, it is more likely to be misinformation and disseminated.

Uncertainty in Message Framing

Uncertainty can also manifest within the message through its framing or word choice. Uncertainty frames are prevalent in misinformation [15,49]. Oh et al [15] illustrated that source ambiguity and content ambiguity are 2 significant features of misinformation. When individuals create a piece of misinformation that lacks evidence and credibility, they often use uncertain words to describe the unreliable source (eg, someone) or the potential rationale (eg, possible, likely) behind the statement. The incorporation of uncertain words can indeed facilitate the spread of misinformation [29,50]. The inclusion of uncertainty expressions in messages leads individuals to perceive the information as more relevant and suitable for themselves [51]. Consequently, if misinformation exhibits a higher level of uncertainty, it is more likely to be accepted and disseminated by the public.

Research Objectives

Our research objective is to explore whether uncertainty features within the information environment can enhance the effectiveness of misinformation detection and spread prediction. To achieve this, we introduce a novel EUP framework specifically designed for both tasks. We seek to assess the standalone effectiveness of the EUP and anticipate that it can augment the capabilities of existing state-of-the-art misinformation detectors and predictors. Therefore, we conducted experiments to answer the following research questions:

- **Research question 1:** Can EUP improve the performances of the state-of-the-art algorithms for misinformation detection and spread prediction?
- **Research question 2:** Can EUP be effective in misinformation detection and spread prediction?

Methods

Overview

Figure 1 offers an overview of the EUP pipeline. The model consists of 4 uncertainty extraction components. Upon receiving a post (denoted as \( p \)), the initial step involves constructing its macro-media environment and micro-communicative environment. This is accomplished by extracting recent news and social media data, respectively. Subsequently, we use a probabilistic model and a similarity calculation method to derive the uncertainty information for the 2 environments mentioned above, denoted as \( I_M \) and \( I_C \). Likewise, we utilized the probabilistic model to capture the uncertainty of the post \( p \) itself, resulting in the representation of message framing denoted as \( I_p \). Simultaneously, the operationalization of uncertainty in the physical environment entails using the number of COVID-19 cases and the volume of news as key indicators, denoted as \( I_r \).

https://ai.jmir.org/2024/1/e47240

JMIR AI 2024 | vol. 3 | e47240 | p.70

(page number not for citation purposes)
Lastly, the 4 vectors are integrated using a gate guided by the extracted post feature \(o\) (which may not necessarily equal \(p\)) from the misinformation detector, such as bidirectional encoder representations from transformers (BERT) \[52\]. The fused vectors \(I\) and \(o\) are then input into the final classifier, typically a multilayer perceptron (MLP), to predict whether \(p\) is fake or real in task 1 and low or high in task 2.

**Figure 1.** An environmental uncertainty perception (EUP) framework for misinformation detection and spread prediction in the COVID-19 pandemic.

### Uncertainty Detection Model

For detecting uncertainty in natural language \[53\], we used a probabilistic model that considers the local n-gram features of sentences. Each n-gram is assigned a weight that reflects its tendency to convey uncertainty. The definition of each feature involves a quadruplet (type, size, context, and aggregation). “Type” signifies the type of n-gram considered, such as lemma or morphosyntactic pattern. “Size” indicates the size of the n-gram. “Context” serves as an indicator, specifying whether the weight is based on the occurrence frequency of the n-gram in an uncertain sentence or on the occurrence frequency of the n-gram as an uncertainty marker. “Aggregation” refers to the method used to consolidate different scores of the n-grams within a sentence. Multimedia Appendix 1 \[49,54-57\] furnishes a summary of the diverse features, denoted as \(F_i\), that are scrutinized in the uncertainty detection model.

Next, we exemplify the calculation of uncertainty using 1 of these features, \(F_1\), as an illustration. \(F_1\) is defined by the quadruplet (Lemma, 1, uncertainty marker, and sum). For each lemma \(w\), we can compute the number of occurrences in the corpus, the number of occurrences in uncertain sentences, and the number of occurrences as an uncertainty marker, denoted as \(F_w\), \(F_{uw}\), and \(F_{um}\) respectively. The conditional probability of a lemma \(w\) becoming an uncertainty marker is calculated using the following equation:

\[
p(c|w) = F_w / F_{w} (1)
\]

where \(c\) represents the class of context uncertainty under analysis, specifically whether it pertains to being an uncertainty marker. Additionally, we introduce a confidence score linked to the probability of mitigating the impact of instances where certain lemmas occur infrequently in the corpus yet yield a high probability:

\[
\text{conf}(w) = 1 - (1 - F_u) (2)
\]

\(F_1\) takes into account both the conditional probability of each lemma \(w\) and the corresponding confidence score in the sentence \(s\), and the formula is calculated as follows:

\[
F^\text{A-Mean}(s) = \text{Mean}(\text{Norm}(\{F_i(s)\}^{\mid F_i \mid=1})) (4)
\]

where \(\text{Norm}(\cdot)\) denotes the normalization.

### Representation of the Macro-Media Environment

We collect news reports from mainstream media outlets released within \(T\) days before the post \(p\) is published to construct a macro-media environment according to the following definition:

\[
M = \{e; e \in E, 0 \leq t_p - t_e \leq T\} (5)
\]

where \(E\) denotes the set of all collected news items, \(M\) denotes the set of news items in the macro-media environment of the post \(p\), and \(t_p\) and \(t_e\) represent the release time of post \(p\) and news \(e\), respectively. For post \(p\) or each news item \(e\), the initial representations are the output of a pretrained language model (eg, BERT \[52\]), denoted as \(p\) and \(e\), respectively.

The macro-media environment is expected to reflect the impact of a trending topic with high uncertainty on the veracity of a post. That is, if a post is related to a trending topic with (vs without) high uncertainty, it is then expected to be more likely misinformation and disseminated. To this end, the representation of the macro-media environment should consider both the correlation between the post and the environment and the
uncertainty of the environment. We first calculate cosine similarity between \( p \) and each news item \( e \) in \( E \):

\[
S(p, e) = \frac{(p-e) \cdot (|p|e)}{||p||e|} \quad (6)
\]

We combine the similarity and environment representations to represent the similarity representation of a post \( p \) to the environment:

\[
\text{Top}(\cdot) \quad (11)
\]

Finally, the macro-media environment of a post \( p \) is represented as an aggregation of the similarity representation of a post \( p \) to the environment (\( S_{e} \)) and the uncertainty representation of the environment (\( U_{\text{env}} \)) using an MLP, denoted as \( I_{\text{U}} \):

\[
I_{\text{U}} = \text{MLP}(S_{e}U_{\text{env}}) \quad (12)
\]

Physical Environment

To measure uncertainty in the physical environment, we collected the daily number of new cases from the start of the COVID-19 outbreak and counted the number of daily news items related to the outbreak, denoted as \( N_{\text{Cases}} \) and \( N_{\text{News}} \), respectively. Intuitively, the higher the number of new cases and news items for a day, the more sensitive the public is to the social environment and the more uncertain the environment is on that day. Thus, the uncertainty factor in the physical environment is defined as follows:

\[
f_{ih} = \text{Norm} \left( \text{log} \left( 1 + \text{abs} \left( \frac{N_{\text{Cases}}}{\text{Cases}_{i - 1}} - \frac{N_{\text{Cases}}}{\text{Cases}_{i - 1}} \right) \right) \times \text{log} \left( 1 + \text{abs} \left( \frac{N_{\text{News}}}{\text{News}_{i - 1}} - \frac{N_{\text{News}}}{\text{News}_{i - 1}} \right) \right) \right) \quad (16)
\]

where \( f_{ih} \) denotes the uncertainty factor at day \( i \) and \( \text{abs} \) is the absolute value operation. For each post, we can obtain the uncertainty factor for its corresponding date \( f_{ih} \).

We added the uncertainty factor of the physical environment to the representations of macro-media environment (\( I_{M} \)), micro-communicative environment (\( I_{C} \)), and post message framing (\( I_{F} \)) to get the representation of the physical environment, denoted as \( I_{P} \):

\[
I_{P} = (f_{ih} \times I_{M}) \quad (17)
\]

Prediction

Prediction With EUP Alone Without Baseline Models

We concatenate the above 4 environment uncertainty features and feed the result into an MLP layer and a softmax layer for the final prediction:

\[
I_{\text{EUP}} = \text{I}_{M} \quad \text{I}_{C} \quad \text{I}_{F} \quad \text{I}_{P} \quad (18)
\]

Prediction With Baseline Models

We expect that our EUP is compatible with and can empower various misinformation detection and prediction algorithms. Therefore, we used an adaptive feature selection approach based on a gate mechanism to accommodate different misinformation detectors:
During training, we minimize the cross-entropy loss.

Ethical Considerations
The study is exempt from ethical review for human subject research for the following reasons. First, the study uses data from 2 publicly available Twitter data sets collected through the official application programming interface (API) of the Twitter platform for gathering tweets. The news data set was obtained from the official websites of news media. Second, the data used in this study are anonymized and do not contain any personally identifiable information. It is also impossible to reidentify individuals from the data set. The data set is stored on a dedicated secure data server, and the analysis is conducted on the platform’s designated site. This process is undertaken for research purposes and adheres to Chinese data privacy laws and regulations. Third, this study does not involve any experimental manipulation of human individuals or other ethical concerns. For instance, it does not include data on children under 18 years of age, which require legally mandated parental or guardian supervision. It also does not encompass sensitive aspects of participants’ behavior or pose any physical, psychological, or economic harm or risk to the research participants.

Data Set and Experiment

Data Set
The statistics and description of our experimental data set are shown in Tables 1 and 2, respectively.

Table 1. Statistics of the data set.a,b

<table>
<thead>
<tr>
<th>Data set</th>
<th>Misinformation detection, n</th>
<th>Spread prediction, n</th>
<th>Total, n</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Real</td>
<td>Fake</td>
<td>Low</td>
</tr>
<tr>
<td>Train</td>
<td>901</td>
<td>1324</td>
<td>1054</td>
</tr>
<tr>
<td>Value</td>
<td>312</td>
<td>430</td>
<td>360</td>
</tr>
<tr>
<td>Test</td>
<td>310</td>
<td>432</td>
<td>358</td>
</tr>
</tbody>
</table>

aNews items in M=58,095. The corresponding mean and range are 988 and 10-2511, respectively.
bTweet items in C=321,656. The corresponding mean and range are 793, 138-1214, respectively.

Table 2. Descriptions of the data set.

<table>
<thead>
<tr>
<th>Data</th>
<th>Features</th>
<th>Size, n</th>
</tr>
</thead>
<tbody>
<tr>
<td>Post</td>
<td>Content, created time, retweet count, veracity label, retweeted label</td>
<td>3709</td>
</tr>
<tr>
<td>News</td>
<td>Content, created time</td>
<td>58,095</td>
</tr>
<tr>
<td>Tweets</td>
<td>Content, created time</td>
<td>321,656</td>
</tr>
</tbody>
</table>

Post
We processed and integrated 2 existing COVID-19 data sets, FibVID [58] and CMU_MisCov19 [59], for our experiments. Both data sets have been labeled for veracity by experts, providing ground-truth labels for our experimental evaluations. For FibVID, we extracted data related to COVID-19, assigning veracity tags as 0 (COVID true) or 1 (COVID fake). We relabeled CMU_MisCov19, classifying calling out or correction, true public health response, and true prevention as real tags, and conspiracy, fake cure, sarcasm or satire, false fact or prevention, fake treatment, and false public health response as fake tags. Furthermore, we used the Twitter API to retrieve the number of retweets for all tweets in both data sets. Subsequently, we categorized the retweet labels as low (when the retweet count is 0) and high (when the retweet count is >0) following an analysis of the distribution of retweet numbers. The data revealed that misinformation was predominantly observed from January to July 2020, coinciding with the period of heightened uncertainty during the pandemic outbreak. Consequently, our focus was directed solely to this specific period, resulting in the extraction of 3709 posts from January to July of 2020.

Macro-Media Environment
We gathered all the news headlines and brief descriptions from the Huffington Post, NPR, and Daily Mail from January to July 2020, as per the methodology outlined previously [12]. Notably, these 3 outlets represent the left-, center-, and right-wing perspectives, contributing to the diversity of news items for our analysis. We then used the keywords “covid,” “coronavirus,” “pneumonia,” “pandemic,” “epidemic,” “infection,” “prevalence,” and “symptom” to filter these data to ensure that the collected data were relevant to COVID-19. We ended up with 58,095 news items from January to July 2020.

Micro-Communicative Environment
We obtained the tweet IDs associated with COVID-19 from an ongoing project [60]. Given the substantial volume, we randomly sampled 1% of these IDs (amounting to approximately 205,581,778 records). Subsequently, using the Twitter API, we...
retrieved the content associated with these IDs, resulting in a
data set comprising 321,656 tweets spanning from January to
July 2020.

Physical Environment
We compiled the daily count of new worldwide COVID-19
cases starting from January 2020, utilizing the Our World in
Data database. Additionally, the daily volume of news data
corresponds to the information we gathered during the same
period.

Experimental Setup

Tasks
We used the proposed model for 2 tasks:

Task 1. Misinformation Detection
The objective was to analyze the text content of a tweet and
ascertain whether it contained misinformation.

Task 2: Spread Prediction
The objective was to evaluate the text content of a tweet to
determine whether it is likely to be retweeted.

Uncertainty Features
Following Jean et al [53], we used WikiWeasel [61], a
comprehensive corpus consisting of paragraphs extracted from
Wikipedia, to compute the frequency of each lemma. The
uncertainty score for each sentence is determined using mean
pooling $F_{A\text{Mean}}$. We leverage [62] to acquire sentence
representations, relying on pretrained BERT models [52] and
subsequent posttraining on news items. In the macro-media
environment and the micro-communicative environment, we
set $T=3$, $r=0.1$, $|C_{\text{min}}|=10$.

Baseline Models
The baseline models considered are listed in Textbox 1.

Textbox 1. Baseline models.

1. **Bidirectional long short-term memory**
   Bidirectional long short-term memory (BiLSTM) [63] is a type of recurrent neural network architecture designed for sequence modeling tasks, particularly in natural language processing. It processes input sequences in both forward and backward directions simultaneously, allowing the model to capture information from both past and future contexts.

2. **Event adversarial neural networks**
   Event adversarial neural networks (EANN) [64] is a model using adversarial training to eliminate event-specific features derived from a convolutional neural network for text (ie, TextCNN).

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

4. **BERT-Emo**
   BERT-Emo [65] is a fake news detection model that integrates multiple sentiment features into BERT.

Evaluation Metrics
For both tasks, we used accuracy and macro-$F_1$-score as
evaluation metrics. Additionally, in task 1, we used $F_1$-scores
for fake ($F_{1\text{fake}}$) and real ($F_{1\text{real}}$), while in task 2, we considered
$F_1$-scores for low ($F_{1\text{low}}$) and high ($F_{1\text{high}}$). Further
implementation details can be found in Multimedia Appendix
1.

Results

Overview

Tables 3 and 4 showcase the performances of the EUP without
baseline models and those of various baseline models, with and
without EUP, for the misinformation detection and spread
prediction tasks, respectively. The results indicate that the
performances of EUP are comparable to those of state-of-the-art
baseline models in both tasks. Moreover, it is noteworthy that
all baseline models exhibit performance improvements when
incorporating EUP for both tasks. These observations suggest
the effectiveness of our proposed EUP.
Table 3. Model performance comparison on the misinformation detection task without the baseline algorithm or without the EUP\textsuperscript{a} module.\textsuperscript{b}

<table>
<thead>
<tr>
<th>Model</th>
<th>Accuracy</th>
<th>Macro-$F_1$-score</th>
<th>$F_1$ fake</th>
<th>$F_1$ real</th>
</tr>
</thead>
<tbody>
<tr>
<td>EUP</td>
<td>0.753</td>
<td>0.739</td>
<td>0.800</td>
<td>0.677</td>
</tr>
<tr>
<td>BiLSTM\textsuperscript{c}</td>
<td>0.733</td>
<td>0.729</td>
<td>0.783</td>
<td>0.683</td>
</tr>
<tr>
<td>BiLSTM + EUP</td>
<td>0.755</td>
<td>0.743</td>
<td>0.798</td>
<td>0.688</td>
</tr>
<tr>
<td>EANN\textsubscript{T}\textsuperscript{d}</td>
<td>0.745</td>
<td>0.730</td>
<td>0.795</td>
<td>0.664</td>
</tr>
<tr>
<td>EANN\textsubscript{T} + EUP</td>
<td>0.767</td>
<td>0.765</td>
<td>0.806</td>
<td>0.708</td>
</tr>
<tr>
<td>BERT\textsuperscript{e}</td>
<td>0.755</td>
<td>0.743</td>
<td>0.797</td>
<td>0.689</td>
</tr>
<tr>
<td>BERT + EUP</td>
<td>0.771</td>
<td>0.767</td>
<td>0.796</td>
<td>0.738</td>
</tr>
<tr>
<td>BERT-Emo</td>
<td>0.749</td>
<td>0.740</td>
<td>0.789</td>
<td>0.691</td>
</tr>
<tr>
<td>BERT-Emo + EUP</td>
<td>0.768</td>
<td>0.763</td>
<td>0.799</td>
<td>0.726</td>
</tr>
</tbody>
</table>

\textsuperscript{a}EUP: Environmental Uncertainty Perception.  
\textsuperscript{b}The best result in each group is in italics.  
\textsuperscript{c}BiLSTM: bidirectional long short-term memory.  
\textsuperscript{d}EANN\textsubscript{T}: event adversarial neural networks.  
\textsuperscript{e}BERT: bidirectional encoder representations from transformers.

Table 4. Model performance comparison on the spread prediction task without the baseline algorithm or without the EUP\textsuperscript{a} module.\textsuperscript{b}

<table>
<thead>
<tr>
<th>Model</th>
<th>Accuracy</th>
<th>Macro-$F_1$-score</th>
<th>$F_1$ low</th>
<th>$F_1$ high</th>
</tr>
</thead>
<tbody>
<tr>
<td>EUP</td>
<td>0.710</td>
<td>0.710</td>
<td>0.719</td>
<td>0.701</td>
</tr>
<tr>
<td>BiLSTM\textsuperscript{c}</td>
<td>0.707</td>
<td>0.705</td>
<td>0.684</td>
<td>0.726</td>
</tr>
<tr>
<td>BiLSTM + EUP</td>
<td>0.734</td>
<td>0.733</td>
<td>0.738</td>
<td>0.729</td>
</tr>
<tr>
<td>EANN\textsubscript{T}\textsuperscript{d}</td>
<td>0.717</td>
<td>0.716</td>
<td>0.734</td>
<td>0.698</td>
</tr>
<tr>
<td>EANN\textsubscript{T} + EUP</td>
<td>0.726</td>
<td>0.726</td>
<td>0.736</td>
<td>0.716</td>
</tr>
<tr>
<td>BERT\textsuperscript{e}</td>
<td>0.728</td>
<td>0.728</td>
<td>0.728</td>
<td>0.728</td>
</tr>
<tr>
<td>BERT + EUP</td>
<td>0.743</td>
<td>0.743</td>
<td>0.752</td>
<td>0.734</td>
</tr>
<tr>
<td>BERT-Emo</td>
<td>0.733</td>
<td>0.733</td>
<td>0.730</td>
<td>0.737</td>
</tr>
<tr>
<td>BERT-Emo + EUP</td>
<td>0.741</td>
<td>0.741</td>
<td>0.733</td>
<td>0.749</td>
</tr>
</tbody>
</table>

\textsuperscript{a}EUP: Environmental Uncertainty Perception.  
\textsuperscript{b}The best result in each group is in italics.  
\textsuperscript{c}BiLSTM: bidirectional long short-term memory.  
\textsuperscript{d}EANN\textsubscript{T}: event adversarial neural networks.  
\textsuperscript{e}BERT: bidirectional encoder representations from transformers.

Ablation Study

We systematically eliminated individual components, namely, macro-media environment, micro-communicative environment, message framing, and physical environment, and assessed the modeling performances on the data set. Tables 5 and 6 illustrate that, under all experimental conditions, performance degrades when any of these components are removed. These results underscore the effectiveness of all 4 uncertainty features of the information environment for both misinformation detection and spread prediction.
Table 5. Ablation study on the misinformation detection task.\(^a\)

<table>
<thead>
<tr>
<th>Model</th>
<th>Accuracy</th>
<th>Macro-$F_1$-score</th>
<th>$F_1$ fake</th>
<th>$F_1$ real</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>EUP(^b)</strong></td>
<td>0.753</td>
<td>0.739</td>
<td>0.800</td>
<td>0.677</td>
</tr>
<tr>
<td>Without $I_M$</td>
<td>0.748</td>
<td>0.738</td>
<td>0.790</td>
<td>0.687</td>
</tr>
<tr>
<td>Without $I_C$</td>
<td>0.745</td>
<td>0.720</td>
<td>0.803</td>
<td>0.637</td>
</tr>
<tr>
<td>Without $I_F$</td>
<td>0.739</td>
<td>0.734</td>
<td>0.778</td>
<td>0.673</td>
</tr>
<tr>
<td>Without $I_P$</td>
<td>0.747</td>
<td>0.730</td>
<td>0.797</td>
<td>0.663</td>
</tr>
<tr>
<td><strong>BiLSTM(^c) + EUP</strong></td>
<td>0.755</td>
<td>0.743</td>
<td>0.798</td>
<td>0.688</td>
</tr>
<tr>
<td>Without $I_M$</td>
<td>0.745</td>
<td>0.741</td>
<td>0.793</td>
<td>0.669</td>
</tr>
<tr>
<td>Without $I_C$</td>
<td>0.741</td>
<td>0.728</td>
<td>0.788</td>
<td>0.668</td>
</tr>
<tr>
<td>Without $I_F$</td>
<td>0.747</td>
<td>0.735</td>
<td>0.791</td>
<td>0.678</td>
</tr>
<tr>
<td>Without $I_P$</td>
<td>0.746</td>
<td>0.742</td>
<td>0.796</td>
<td>0.665</td>
</tr>
<tr>
<td><strong>BERT(^d) + EUP</strong></td>
<td>0.771</td>
<td>0.767</td>
<td>0.796</td>
<td>0.738</td>
</tr>
<tr>
<td>Without $I_M$</td>
<td>0.762</td>
<td>0.754</td>
<td>0.801</td>
<td>0.707</td>
</tr>
<tr>
<td>Without $I_C$</td>
<td>0.764</td>
<td>0.761</td>
<td>0.807</td>
<td>0.696</td>
</tr>
<tr>
<td>Without $I_F$</td>
<td>0.761</td>
<td>0.752</td>
<td>0.800</td>
<td>0.705</td>
</tr>
<tr>
<td>Without $I_P$</td>
<td>0.758</td>
<td>0.751</td>
<td>0.795</td>
<td>0.707</td>
</tr>
</tbody>
</table>

\(^a\)The best result in each group is in italics.
\(^b\)EUP: Environmental Uncertainty Perception.
\(^c\)BiLSTM: bidirectional long short-term memory.
\(^d\)BERT: bidirectional encoder representations from transformers.
Table 6. Ablation study on the spread prediction task.\textsuperscript{a}

<table>
<thead>
<tr>
<th>Model</th>
<th>Accuracy</th>
<th>Macro-$F_1$-score</th>
<th>$F_{1 \text{ low}}$</th>
<th>$F_{1 \text{ high}}$</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>EUP</strong>\textsuperscript{b}</td>
<td>0.710</td>
<td>0.710</td>
<td>0.719</td>
<td>0.701</td>
</tr>
<tr>
<td>Without $I_M$</td>
<td>0.697</td>
<td>0.696</td>
<td>0.715</td>
<td>0.676</td>
</tr>
<tr>
<td>Without $I_C$</td>
<td>0.695</td>
<td>0.694</td>
<td>0.712</td>
<td>0.677</td>
</tr>
<tr>
<td>Without $I_F$</td>
<td>0.702</td>
<td>0.702</td>
<td>0.714</td>
<td>0.689</td>
</tr>
<tr>
<td>Without $I_p$</td>
<td>0.708</td>
<td>0.707</td>
<td>0.721</td>
<td>0.692</td>
</tr>
<tr>
<td><strong>BiLSTM\textsuperscript{c} + EUP</strong></td>
<td>0.734</td>
<td>0.733</td>
<td>0.738</td>
<td>0.729</td>
</tr>
<tr>
<td>Without $I_M$</td>
<td>0.724</td>
<td>0.723</td>
<td>0.735</td>
<td>0.711</td>
</tr>
<tr>
<td>Without $I_C$</td>
<td>0.721</td>
<td>0.721</td>
<td>0.716</td>
<td>0.726</td>
</tr>
<tr>
<td>Without $I_F$</td>
<td>0.717</td>
<td>0.716</td>
<td>0.731</td>
<td>0.702</td>
</tr>
<tr>
<td>Without $I_p$</td>
<td>0.726</td>
<td>0.723</td>
<td>0.753</td>
<td>0.693</td>
</tr>
<tr>
<td><strong>BERT\textsuperscript{d} + EUP</strong></td>
<td>0.743</td>
<td>0.743</td>
<td>0.752</td>
<td>0.734</td>
</tr>
<tr>
<td>Without $I_M$</td>
<td>0.741</td>
<td>0.739</td>
<td>0.764</td>
<td>0.713</td>
</tr>
<tr>
<td>Without $I_C$</td>
<td>0.741</td>
<td>0.738</td>
<td>0.766</td>
<td>0.711</td>
</tr>
<tr>
<td>Without $I_F$</td>
<td>0.736</td>
<td>0.735</td>
<td>0.753</td>
<td>0.716</td>
</tr>
<tr>
<td>Without $I_p$</td>
<td>0.740</td>
<td>0.738</td>
<td>0.759</td>
<td>0.717</td>
</tr>
</tbody>
</table>

\textsuperscript{a}The best result in each group is in italics.
\textsuperscript{b}EUP: Environmental Uncertainty Perception.
\textsuperscript{c}BiLSTM: bidirectional long short-term memory.
\textsuperscript{d}BERT: bidirectional encoder representations from transformers.

The Effect of the Day Parameter $T$

To explore the impact of the day parameter ($T$) on the results during the construction of the macro-media environment and the micro-communicative environment, we experimented with different values of $T$. Specifically, we sequentially set $T=1, 3, 5, 7, 9$ for the BERT + EUP model, and the experimental results are depicted in Figure 2. Despite the fact that increasing $T$ results in larger macro-media and micro-communicative environments, the optimal performance was achieved when $T=1$.

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

![Figure 2](https://ai.jmir.org/2024/1/e47240)
The Effect of the Rate Parameter $r$

We maintained the setting $T=3$ and systematically varied $r$, using values of 0.05, 0.1, 0.15, 0.2, 0.25, and 0.3 on the BERT + EUP model to examine the impact of $r$ on the experimental results, as illustrated in Figure 3. The accuracy performance exhibited fluctuations with varying values of $r$. Notably, the highest accuracy for both tasks was observed when $r=0.1$.

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

![Figure 3](image)

Evaluation on Imbalanced Data

In real-world scenarios, the distribution of real and fake information often exhibits significant imbalance. To evaluate the efficacy of our proposed EUP framework on unbalanced data sets, we conducted tests on data sets with varying ratios of real to fake data, ranging from 10:1 to 100:1. We measured and reported macro-$F_1$-scores and standardized partial area under the curve (AUC) with a false-positive rate of at most 0.1 (i.e., spAUCFPR≤0.1 [66]) to assess the effectiveness of our EUP framework in handling nonbalanced data sets. As depicted in Figure 4, EUP yields relative improvements of 21.5% and 5.7% in macro-$F_1$-score and spAUCFPR≤0.1, demonstrating its effectiveness on unbalanced data sets.

**Figure 4.** Performance of macroF1 and spAUC values across datasets with varying ratios.

![Figure 4](image)

Discussion

**Principal Findings**

First, this study enhances scholars’ comprehension of the misinformation detection and spread prediction problem by highlighting the significance of uncertainty in information environments. Notably, this research contributes to the literature by recognizing uncertainty features in the information environments of misinformation as a pivotal factor for improving detection and prediction algorithms during a pandemic. Our findings underscore that the EUP alone is sufficient for both tasks and has the potential to enhance the capabilities of state-of-the-art algorithms. In contrast to prior misinformation research that primarily concentrates on post content (such as post theme, sentiments, and linguistic characteristics, as seen in [6,11,29]) and network connections (e.g., number of followers [25]) on social media, this study advances scholars’ understanding of the misinformation problem by emphasizing the importance of uncertainty in information environments.
environments. Recognizing and incorporating uncertainty as a fundamental concept in misinformation detection and spread prediction during crises hold theoretical significance. This is particularly relevant as a crisis is characterized by its unpredictable, unexpected, and nonroutine nature, inherently giving rise to uncertainty [38,67]. This uncertainty has been theorized to compel individuals to seek information as a coping mechanism for dealing with the anxiety and pressure generated by uncertainty. This process allows people to diminish uncertainty, restore a sense of normalcy, and alleviate anxiety [14,68]. Regrettably, this coping mechanism can inadvertently fuel the proliferation and dissemination of misinformation, particularly when there is a lack of timely and accurate information, contributing to the concurrent occurrence of an infodemic [6,11,50]. The current research seeks to advance the literature by establishing the legitimacy of uncertainty in the information environments of misinformation as a central indicator for the detection and prediction of misinformation during public health crises.

Second, this study delves into the intricacies of uncertain information environments for misinformation across 4 distinct scales, namely, the physical environment, macro-media environment, micro-communicative environment, and message framing. Our findings demonstrate the effectiveness of all 4 uncertainty features in misinformation detection and spread prediction. In contrast to prior misinformation literature during the COVID-19 pandemic, which often overlooked the role of the information environment in increasing the likelihood of misinformation dissemination, our research emphasizes the importance of considering uncertainty beyond the content of misinformation itself, such as ambiguous wording [29,50]. Our study broadens the concept of linguistic uncertainty in misinformation message framing to encompass a more comprehensive uncertainty across various information environments. We define uncertainty in information environments using a multiscale approach that highlights the significance of the interaction between the physical environment and macro-/micro-media environments. This approach diverges from focusing on a single dimension, such as ambiguities about official guidelines and news reports [18], or the misinformation framing strategy on social media [29].

Third, our findings indicate that uncertainties in information environments play a crucial role as motivators for the emergence and spread of misinformation. While previous studies have provided preliminary evidence suggesting that uncertainty stemming from government policies and news media could coincide with the occurrence of related misinformation during the COVID-19 pandemic, often relying on descriptive big data analyses [3,32], our study contributes stronger empirical evidence. We leverage machine learning techniques to demonstrate that uncertainty arising from the crisis and crisis communication through media can indeed incentivize individuals to generate and disseminate misinformation. Significantly, our findings revealed that the algorithm achieved its best performance for both detection and spread prediction tasks when incorporating items from the information environments published 1 day before the post (T=1). This discovery emphasizes the acute impact of uncertainty in the information environment on the emergence and spread of misinformation, underscoring the importance of timely uncertainty reduction in crisis communication. Furthermore, the algorithm attained the highest accuracies when it included items highly relevant to the post but with an appropriate size (r=0.1). This rationale is reasonable, as a too-small r may fail to encompass enough misinformation-related items, while a larger r might include a significant amount of irrelevant information. The evidence theoretically establishes a connection between crisis communication research and misinformation research, reinforcing the notion that crisis communication and misinformation containment are 2 intertwined aspects of crisis management [3].

This study offers significant practical implications for misinformation detection and spread prediction. First, unlike previous studies that separately investigated computational frameworks for these tasks [24,29], this study introduces a unified uncertainty–based framework capable of addressing both tasks simultaneously. Second, our framework operates instantaneously, as it only requires easily accessible data such as posts, mainstream news, and relevant social media discussions published a few days prior. Moreover, the uncertainty detection algorithm has been trained using external data, rendering our algorithm easy to implement and capable of providing timely detection and prediction for streaming textual data. Third, this study affirms the effectiveness of uncertainty in various information environments for detecting and predicting misinformation on social media. Hence, the 4 proposed uncertainty components in information environments could be leveraged by social media platforms to improve the accuracy of misinformation detection and spread prediction, thereby safeguarding individuals from harm caused by infodemic. The benefits offered by our algorithm may serve as an impetus for integrating uncertainty components into practical systems.

Limitations and Future Work
This study is the first to incorporate the uncertainty present in the information environment of a post for both misinformation detection and spread prediction. However, it has some limitations. First, our framework concentrated solely on text-only detection and prediction. Future work should extend the framework to incorporate multimodal and social graph–based detection. Second, we used an uncertainty detection algorithm developed from a generic corpus sourced from Wikipedia. Nevertheless, past research has indicated that expressions of uncertainty may vary slightly across domains [53]. In other words, uncertainty expressions in the context of the COVID-19 pandemic may differ from those in general situations. Therefore, future work should aim to enhance our uncertainty measure by utilizing a corpus specifically designed for uncertainty detection in the discourse related to COVID-19.

Conclusions
We introduced an EUP framework for both misinformation detection and spread prediction. Our framework delves into uncertainty within information environments across 4 scales: the physical environment, macro-media environment, micro-communicative environment, and message framing. The experiments demonstrated the effectiveness of our proposed
uncertainty components in enhancing the performance of existing models. There are several directions for further investigation and extension of this work. First, we can explore the impact of different news and social media environments (e.g., biased vs neutral; left wing vs right wing) on the emergence and spread of misinformation. Second, extending our algorithms to include multimodal misinformation detection could be beneficial, as misinformation increasingly incorporates images and videos. Third, investigating the interaction between misinformation detection and spread prediction using a multitask, transfer-learning model is a promising avenue, given the shared uncertainty framework identified in this study for both tasks.

Acknowledgments
This study was supported by Open Funding Project of the State Key Laboratory of Communication Content Cognition (grant number 20G01).

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

Conflicts of Interest
None declared.

Multimedia Appendix 1
Uncertainty features.

References


Abbreviations

API: application programming interface
AUC: area under the curve
BERT: bidirectional encoder representations from transformers
BiLSTM: bidirectional long short-term memory
EANN: event adversarial neural networks
EUP: Environmental Uncertainty Perception
MLP: multilayer perceptron
spAUCFPR: standardized partial area under the curve with a false-positive rate
TextCNN: convolutional neural network for text
Original Paper

Identifying Frailty in Older Adults Receiving Home Care Assessment Using Machine Learning: Longitudinal Observational Study on the Role of Classifier, Feature Selection, and Sample Size

Cheng Pan\(^1\), MPhil; Hao Luo\(^2\), PhD; Gary Cheung\(^3\), PhD; Huiquan Zhou\(^2\), PhD; Reynold Cheng\(^1\), PhD; Sarah Cullum\(^3\), PhD; Chuan Wu\(^1\), PhD

\(^1\)Department of Computer Science, The University of Hong Kong, Hong Kong, China (Hong Kong)
\(^2\)Department of Social Work and Social Administration, The University of Hong Kong, Hong Kong, China (Hong Kong)
\(^3\)Department of Psychological Medicine, School of Medicine, The University of Auckland, Auckland, New Zealand

Corresponding Author:
Hao Luo, PhD
Department of Social Work and Social Administration
The University of Hong Kong
CJT 521, Jockey Club Tower
Pokfulam Road
Hong Kong
China (Hong Kong)
Phone: 852 68421252
Email: haoluo@hku.hk

Abstract

Background: Machine learning techniques are starting to be used in various health care data sets to identify frail persons who may benefit from interventions. However, evidence about the performance of machine learning techniques compared to conventional regression is mixed. It is also unclear what methodological and database factors are associated with performance.

Objective: This study aimed to compare the mortality prediction accuracy of various machine learning classifiers for identifying frail older adults in different scenarios.

Methods: We used deidentified data collected from older adults (65 years of age and older) assessed with interRAI-Home Care instrument in New Zealand between January 1, 2012, and December 31, 2016. A total of 138 interRAI assessment items were used to predict 6-month and 12-month mortality, using 3 machine learning classifiers (random forest [RF], extreme gradient boosting [XGBoost], and multilayer perceptron [MLP]) and regularized logistic regression. We conducted a simulation study comparing the performance of machine learning models with logistic regression and interRAI Home Care Frailty Scale and examined the effects of sample sizes, the number of features, and train-test split ratios.

Results: A total of 95,042 older adults (median age 82.66 years, IQR 77.92-88.76; n=37,462, 39.42% male) receiving home care were analyzed. The average area under the curve (AUC) and sensitivities of 6-month mortality prediction showed that machine learning classifiers did not outperform regularized logistic regressions. In terms of AUC, regularized logistic regression had better performance than XGBoost, MLP, and RF when the number of features was ≤80 and the sample size ≤16,000; MLP outperformed regularized logistic regression in terms of sensitivities when the number of features was ≥240 and the sample size ≥4000. Conversely, RF and XGBoost demonstrated higher specificities than regularized logistic regression in all scenarios.

Conclusions: The study revealed that machine learning models exhibited significant variation in prediction performance when evaluated using different metrics. Regularized logistic regression was an effective model for identifying frail older adults receiving home care, as indicated by the AUC, particularly when the number of features and sample sizes were not excessively large. Conversely, MLP displayed superior sensitivity, while RF exhibited superior specificity when the number of features and sample sizes were large.

(JMIR AI 2024;3:e44185) doi:10.2196/44185
KEYWORDS

machine learning; logistic regression; frailty; older adults; home care; sample size; features; data set; model; home care; mortality prediction; assessment

Introduction

Frailty is a syndrome characterized by an increased vulnerability to adverse health outcomes, including falling, hospitalization, physical decline, and mortality [1]. Frailty should be detected as early as possible since it is potentially preventable and treatable [2]. In community settings, timely identification of frailty allows the implementation of early interventions that could reduce care costs and improve the “ability of older persons to age in place” [3]. In clinical and long-term care settings, identifying frail older adults could facilitate more individualized and tailored health care planning [4,5]. Therefore, efficient and accurate clinical tools are pivotal to the early identification of frailty among at-risk older adults.

Numerous methods have been applied to measure frailty. A recent systematic review identified 21 conceptual definitions and 59 operational definitions of frailty from 68 studies [6]. This review concluded that definitions of frailty can be classified into 3 categories focusing on different dimensions. The first is represented by the Cardiovascular Health Study (CHS) Index based on Fried’s “frailty phenotype” model, which focuses on the physical dimensions of frailty [7-10]. The second category is represented by the Frailty Index, originally proposed by Rockwood and Mitnitski [11,12], which considers frailty as a syndrome capturing the accumulative gradient of deficits. This category of definitions covers other dimensions of frailty, including cognitive, psychological, nutritional, and social factors [11,13]. The third category considers the social dimension of frailty, which has a significant relationship with undesirable adverse health outcomes [14-16]. Despite differences in theoretical frameworks adopted by different frailty measures, existing frailty indices are typically constructed by summing up the number of deficits or scores of assessment items using equal weighting. Arguably, different deficits from various domains may impact overall frailty status differently, and these differences should be considered when measuring frailty. In addition to accounting for the multifactorial nature of frailty, a successful definition of frailty [12] must demonstrate satisfactory criterion validity. Since frailty is noncontroversially linked with vulnerability, a valid measure of frailty must accurately predict adverse outcomes, such as death, institutionalization, hospitalization, physical decline, and falls. Mortality is the most objective measure that is less susceptible to measurement error and, thus, is the most widely used outcome for assessing the predictive validity of frailty measures [9,17-20].

 Routinely collected data from health information systems have become increasingly available in recent years, and clinical big data analytics featured by machine learning techniques are ever-evolving [21-23]. In contrast to conventional regression approaches, classifiers used in machine learning, such as random forest (RF), support vector machines, and neural networks, have the advantages of learning and generating predictions by examining large-scale databases of complex clinical information [18,20,24-26]. Therefore, it is reasonable to hypothesize that applying machine learning techniques to large-scale data collected from health information systems can improve the accuracy of mortality prediction for identifying frail older persons who may benefit from early interventions. However, the literature remains unclear whether machine learning techniques can outperform conventional regression models in identifying frail older adults [18,19,27].

In this study, we used routinely collected health information of people receiving home care in New Zealand from interRAI-Home Care (interRAI-HC) assessment to examine the performance of various machine learning classifiers in mortality prediction for identifying frailty. In this study, we conducted a simulation study to address the following research questions: (1) does the performance of machine learning models exceed that of the interRAI-HC Frailty Scale, which was developed using conventional regression models [28], in identifying frailty? (2) what are the performances of different machine learning models? and (3) what are the effects of sample size, number of features, and the ratio of training to test data on predictive accuracy?

Methods

Data Source and Participants

In this retrospective observational study, we used deidentified health information routinely collected from older adults assessed using the interRAI-HC assessment (version 9.1). The interRAI-HC assessment was developed by a network of health researchers in over 35 countries [29]. interRAI assessments are mandatory in aged residential care and home and community services for older people living in the community in New Zealand. Our participants were from all 20 District Health Boards in New Zealand and included all community-dwelling older adults who were receiving public-funded home care or assessed for long-term aged residential care. Trained interRAI assessors collect comprehensive health information on older adults, including their demographic, clinical, psychosocial, and functional details. The interRAI-HC assessment embeds over 100 potential deficits of older adults that can be used to identify frailty. Table S1 in Multimedia Appendix 1 summarizes the variables used for identifying frail older adults. Ethnicity was not included to increase generalizability beyond New Zealand.

We included adults 65 years of age or older for whom at least 1 interRAI-HC assessment had been completed between January 1, 2012, and December 31, 2016. Only the most recent interRAI-HC assessment (defined as the index assessment) of each individual within this period was used in the analysis and the date of the most recent assessment was defined as the index date. The individuals were followed from the index date until the date of death or December 31, 2019, whichever came first.

Ethical Considerations

The University of Auckland Human Participant Ethics Committee provided ethics approval for this study (023801).
Measures

Outcomes

Outcomes of interest were 6-month and 12-month mortality. Mortality data were retrieved from the Ministry of Health Mortality Dataset that contains information of all registered deaths in New Zealand. These two-time points were chosen because (1) older adults receiving home care are associated with a higher risk of mortality and shorter survival compared with their counterparts who are not receiving home care and (2) these are outcomes commonly used in previous studies examining the association between frailty and mortality [30-33] and few previous studies using interRAI data [34-36].

Features Used in Machine Learning Models

Features of interest included 138 interRAI-HC assessment items covering 11 broad domains, demographics, cognition, communication and vision, mood and behavior, psychosocial well-being, functional status, continence, disease diagnoses, health conditions, oral and nutrition status, and skin conditions. Table S1 in Multimedia Appendix 1 presents the details of features used to identify frail older individuals.

Assessment items that had a missing percentage of over 10% were excluded from this study. Multiple interRAI-HC assessment variables with a response indicating that the activity did not occur during the assessment were considered missing, and the missing data imputation was implemented for these responses.

Established Frailty Scales (Benchmark)

The interRAI-HC Frailty Scale was used as the benchmark for evaluating the predictive performance of machine learning algorithms. The interRAI-HC Frailty Scale was developed and validated using assessments collected from multiple and diverse countries worldwide [28]. Table S2 in Multimedia Appendix 1 summarizes the variables used in constructing the interRAI-HC Frailty Scale.

Machine Learning and Logistic Regression Models

We applied 3 state-of-the-art machine learning models and regularized logistic regression to predict 6-month and 12-month mortality using the features available from interRAI-HC. The RF is a machine learning algorithm that uses decision trees [37]. The RF provides highly accurate predictions with a very large number of input variables [38]. The eXtreme Gradient Boosting (XGBoost) is an optimized algorithm designed to implement parallel tree boosting that can predict results extremely efficiently and accurately based on its scalability and efficiency in all scenarios [39]. Multilayer perceptron (MLP) is one of the most popular paradigms of artificial neural networks. MLP decreases the output error by adjusting the weights of predictive variables through an iterative learning process [40].

Regularized logistic regression is a variant of logistic regression using regularization to prevent overfitting and improve the performance of logistic regression. Two popular types of regularized logistic regressions are Least Absolute Shrinkage and Selection Operator (LASSO) regularization with the L1 penalty [41] and Ridge regularization with the L2 penalty [42].

In this study, we implemented hyperparameter tuning to regularize logistic regression (hereafter referred to as logistic regression), RF, MLP, and XGBoost by performing a randomized grid search using all home care (HC) assessment items. The best hyperparameters for each classifier were determined by 10-fold cross-validation (Table S5 in Multimedia Appendix 1). We used iterative imputation [43] to handle the missing values and the default threshold of 0.5 was used in training [27]. We conducted a sensitivity analysis to compare the performance of the models with and without imputation in selected conditions, that is, only the minimum and maximum sample sizes and the number of features were selected for comparison due to the expensive computation power required.

The preliminary results suggested that our data are imbalanced, as the majority of individuals survived within 6 or 12 months. We therefore rebalanced the training data (but not the test data) using random oversampling [44], while keeping the test data unchanged. Our primary findings are presented with the results obtained after rebalancing the data. The results using the original imbalanced data set can be found in Multimedia Appendix 1. Specifically, to initiate the hyperparameter tuning process, we performed hyperparameter tuning using grid search. For each combination of hyperparameters, within each iteration of the 10-fold cross-validation loop, we applied oversampling to the training set, and the model was trained on the oversampled training set using the current combination of hyperparameters. The model’s performance was evaluated on the validation set. After all combinations of hyperparameters have been evaluated, we selected the combination that gave the best average performance. The process of data preprocessing, training, prediction, and evaluation is illustrated in Figure 1.
Simulation Design

We conducted a Monte Carlo simulation to compare the performance of different machine learning methods and logistic regression under different experimental conditions, characterized by different sample sizes, the number of features, and training test split ratios. There were 72 experimental conditions for each model (4 sample sizes, 6 feature numbers, and 3 training test split ratios). Each of these conditions was repeated 1000 times to assess their variability. We used sample sizes equaling 1000, 4000, 16,000, and 95,042; the number of features equaling 10, 20, 30, 40, 80, and 138; and training test split ratios equaling 7:3, 8:2, and 9:1 in our simulation. We selected these sample sizes and feature numbers because they are commonly encountered in existing studies on frailty measurement [17,19,45-48] and are values that are testable using the current database. The training split ratios are widely used in studies using machine learning [18,27,36,49,50]. We chose a limited number under each domain to keep the simulations to a manageable scale.

Evaluation of Model Performance

We randomly split the data into a training sample and a test sample with different training test ratios. We evaluated model performances using the test sample. The discrimination ability of each classifier was measured by the area under the curve (AUC) [51], sensitivity, (also referred to as the true positive rate), and specificity (also known as the true negative rate) as the primary criteria because these are criteria widely accepted by the clinicians. Since frailty is reversible and may be attenuated by noninvasive interventions such as exercise, reduction of polypharmacy, and adequate nutrition [52], high sensitivity is viewed as more important than high specificity in this context if a trade-off needs to be made. F1-score [53], accuracy and precision (also called positive predictive value) [47,54,55] were also constructed and assessed to allow comparisons with studies that reported only these outcomes. Note, that as each experimental condition was repeated 1000 times to address the potential impact of randomization, we computed the mean and SDs of all performance indices across 1000 replications. The 95% CI for the performance metrics was computed from 1000 runs for each scenario.

Results

We included 95,042 older adults after excluding 4676 individuals who were younger than 65 years of age and 51 individuals with incorrect records (eg, the date of death was earlier than the assessment date, invalid date of birth, or an incorrect assessment date). Table 1 summarizes the characteristics of study subjects, stratified by whether the person died within 6 months. About half of the subjects were aged between 80 and 89 years (80-84 years: n=21,947, 23.09%; 85-89 years: n=23,906, 25.15%). Women accounted for 57,580 (60.58%) of the sample, and 83,590 (87.95%) were European.
A total of 12,401 (13.05%) subjects died within 6 months following the index assessment. Table S19 in Multimedia Appendix 1 documents the characteristics of the study subjects, stratified by whether the person died within 1 year.

Table S4 in Multimedia Appendix 1 presents the results of the sensitivity analysis comparing the performance of the models with and without imputation. The findings suggest that the data imputation was necessary as the imputed data set outperformed the unimputed data set in most of the conditions tested.

After comparing the performance of penalty terms none, L1, and L2, the LASSO regression regularization (L1) and Ridge regularization (L2) were used in 6-month and 12-month mortality prediction, respectively. We compared the average AUC of each classifier as the number of features increased for 6-month mortality prediction (Figure 2). Overall, the performance of all methods improved considerably as the number of features increased. Specifically, in most scenarios, when the number of features increased to 30, four classifiers demonstrated significantly higher AUC than the interRAI-HC Frailty Scale. LASSO regression generally demonstrated higher or comparable AUC than RF, MLP, and XGBoost. However, in the specific scenario where the sample size was 95,042 and the number of features was 40 or less, MLP showed a slightly better average AUC than LASSO regression. In addition, when the sample size was 95,042, and the number of features increased to 138, XGBoost achieved the highest average AUC of 0.79 (95% CI 0.79-0.80).

Figure 3 shows the average sensitivities across all experimental conditions. The 3 machine learning classifiers and LASSO regression had lower sensitivities than the interRAI-HC frailty scale when the sample size was 1000. As the sample size increased to 4000 and the number of features increased to 20, MLP and LASSO regression outperformed the benchmark scale with the highest average sensitivity of 0.77 (95% CI 0.72-0.79) observed in MLP when the sample size was 95,042, and the number of features was 138. Meanwhile, all classifiers demonstrated higher average specificities than the interRAI-HC Frailty Scale in all scenarios (Figure 4). The RF and XGBoost demonstrated higher specificities than LASSO regression, with RF achieving the highest average specificities of 0.98 (95% CI 0.98-0.98) when the sample size was 95,042 and the number of features was 138.

Based on the simulation results, it was observed that the test size ratios did not have a significant impact on the average AUC, sensitivities, and specificities, as shown in Figure 5. The 12-month and 6-month mortality predictions were comparable (Figures S1-S4 in Multimedia Appendix 1). However, the overall performance of logistic regression on the 12-month mortality prediction was worse than the 6-month prediction. Compared to the 6-month mortality prediction, machine learning classifiers performed slightly better average sensitivities and worse average AUCs and specificities on 12-month mortality prediction. Tables S5-S18 and S20-S33 in Multimedia Appendix 1 summarize AUC, sensitivity, specificity, $F_1$-score, accuracy, and precision.

Our simulation was also conducted on the imbalanced data set, and we observed a similar result in terms of average AUCs. Regularized logistic regression had a higher AUC than XGBoost, MLP, and RF, especially when the number of features was less than or equal to 80 and the sample size was less than or equal to 16,000. However, as the number of features and sample sizes increased, XGBoost slightly outperformed regularized logistic regression. In terms of sensitivities, regularized logistic regression significantly outperformed machine learning classifiers in all scenarios, while machine learning classifiers had higher specificities than regularized logistic regression in all scenarios. Additionally, the findings for 12-month and 6-month mortality prediction were similar. However, machine learning classifiers performed slightly better in average sensitivities, but worse in average AUCs and specificities for 12-month mortality prediction compared to 6-month mortality prediction. Multimedia Appendix 1 has been included to summarize the results of the imbalanced data set (Tables S34-S62 and Figures S9-S12 in Multimedia Appendix 1).
Table 1. Sample characteristics of 6-month mortality.

<table>
<thead>
<tr>
<th>Characteristics</th>
<th>HC\textsuperscript{a} (N=95,042)</th>
<th>6-month deceased (n=12,401)</th>
<th>6-month survived (n=82,641)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Age (years)</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>65-69, n (%)</td>
<td>5906 (6.21)</td>
<td>693 (5.59)</td>
<td>5213 (6.31)</td>
</tr>
<tr>
<td>70-74, n (%)</td>
<td>9623 (10.12)</td>
<td>1065 (8.59)</td>
<td>8558 (10.36)</td>
</tr>
<tr>
<td>75-79, n (%)</td>
<td>15,284 (16.08)</td>
<td>1770 (14.27)</td>
<td>13,514 (16.35)</td>
</tr>
<tr>
<td>80-84, n (%)</td>
<td>21,947 (23.09)</td>
<td>2662 (21.47)</td>
<td>19,285 (23.34)</td>
</tr>
<tr>
<td>85-89, n (%)</td>
<td>23,906 (25.15)</td>
<td>3312 (26.71)</td>
<td>20,594 (24.92)</td>
</tr>
<tr>
<td>90-94, n (%)</td>
<td>14,370 (15.12)</td>
<td>2160 (17.42)</td>
<td>12,210 (14.77)</td>
</tr>
<tr>
<td>≥100, n (%)</td>
<td>412 (0.43)</td>
<td>85 (0.69)</td>
<td>327 (0.40)</td>
</tr>
<tr>
<td>Mean (SD)</td>
<td>82.66 (7.61)</td>
<td>83.59 (7.71)</td>
<td>82.52 (7.59)</td>
</tr>
<tr>
<td><strong>Gender, n (%)</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Female</td>
<td>57,580 (60.58)</td>
<td>6362 (51.30)</td>
<td>51,218 (61.98)</td>
</tr>
<tr>
<td>Male</td>
<td>37,462 (39.42)</td>
<td>6039 (48.70)</td>
<td>31,423 (38.02)</td>
</tr>
<tr>
<td><strong>Ethnicity, n (%)</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>European</td>
<td>83,590 (87.95)</td>
<td>11,128 (89.73)</td>
<td>72,462 (87.68)</td>
</tr>
<tr>
<td>Maori</td>
<td>5321 (5.60)</td>
<td>730 (5.89)</td>
<td>4591 (5.56)</td>
</tr>
<tr>
<td>Pacific Island</td>
<td>2948 (3.10)</td>
<td>267 (2.15)</td>
<td>2681 (3.24)</td>
</tr>
<tr>
<td>Asian</td>
<td>2304 (2.42)</td>
<td>197 (1.59)</td>
<td>2107 (2.55)</td>
</tr>
<tr>
<td>Middle eastern or Latin American or African</td>
<td>352 (0.37)</td>
<td>25 (0.20)</td>
<td>327 (0.40)</td>
</tr>
<tr>
<td>Other ethnicity</td>
<td>527 (0.55)</td>
<td>54 (0.44)</td>
<td>473 (0.57)</td>
</tr>
<tr>
<td><strong>Marital status, n (%)</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Married or civil union or de facto</td>
<td>82,401 (86.70)</td>
<td>10,936 (88.19)</td>
<td>71,465 (86.48)</td>
</tr>
<tr>
<td>Never married</td>
<td>4486 (4.72)</td>
<td>539 (4.35)</td>
<td>3947 (4.78)</td>
</tr>
<tr>
<td>Widowed</td>
<td>2116 (2.23)</td>
<td>240 (1.94)</td>
<td>1876 (2.27)</td>
</tr>
<tr>
<td>Separated or divorced</td>
<td>5999 (6.31)</td>
<td>683 (5.51)</td>
<td>5316 (6.43)</td>
</tr>
<tr>
<td>Others</td>
<td>40 (0.04)</td>
<td>3 (0.02)</td>
<td>37 (0.04)</td>
</tr>
</tbody>
</table>

\textsuperscript{a}HC: home care.

Figure 2. Average AUCs of classifiers and frailty scale for 6-month mortality prediction on balanced data set. AUC: area under the curve; HC: home care; LR: logistic regression; MLP: multilayer perceptron; XGBoost: extreme gradient boosting.
Figure 3. Average sensitivities of classifiers and frailty scale for 6-month mortality prediction on balanced data set. HC: home care; LR: logistic regression; MLP: multilayer perceptron; XGBoost: extreme gradient boosting.

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

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

Principal Findings

In this retrospective study of older adults with the mandated standardized interRAI-HC assessment in New Zealand, we performed a series of simulations to evaluate the role of machine learning classifiers, features, and sample sizes on mortality prediction in identifying frail older individuals. We found that in most scenarios, particularly when dealing with large sample sizes and large numbers of features, 4 classifiers demonstrated significantly higher AUCs and sensitivities compared to the interRAI-HC Frailty Scale. All classifiers showed higher average specificities than the interRAI-HC Frailty Scale across all scenarios. Our simulation results showed that the predictive performance differed significantly by using different numbers of randomly selected features, varied sample sizes, and performance measures. Compared to machine learning classifiers, that is, RF, MLP, and XGBoost, logistic regressions provided higher average AUCs on 6-month mortality prediction when the number of features and sample sizes were not excessive. Even with a high number of features and very large samples, only slight improvements in average AUCs were observed in MLP and XGBoost. However, when the number of features and sample sizes were large, MLP demonstrated superior sensitivity, whereas RF exhibited superior specificity.

Interpretation in the Light of the Published Literature

In recent years, machine learning techniques have started to be used in various large-scale health care data sets to develop predictive algorithms for various adverse health outcomes, including hospitalization, mortality, and frailty in different populations [18,20,24,56]. For example, a recent study showed that by using only 10 or 11 features and 592 study subjects, the machine learning classifier support vector machines identified frail older adults with over 75% accuracy [45]. Another study also showed that by using 16 features, the machine learning classifier gradient boosting achieved 90% AUC on 30-day mortality prediction in patients with heart failure [19]. However, due to limitations in sample size and the number of available features, no study has systematically examined the role of methodological and database factors in the performance of various machine learning techniques. To our knowledge, our study is the first to use high-quality health care data of older adults receiving home care to investigate the performance of machine learning classifiers in identifying frail persons compared to an existing clinical scale and conventional logistic regressions. It is also the first to elucidate to what extent the performance is associated with the choice of classifier, sample size, and the number of features.

Contrary to our hypothesis, the application of machine learning classifiers did not improve the performance of mortality prediction for identifying frail older adults, as evaluated by AUC. This finding indicates that regularized logistic regression can perform sufficiently well and save computational resources when a well-structured, high-quality data source is used. One possible explanation for this result could be the nature of the features, as most of the items used to identify frail older adults are binary. Another reason may be the high reliability of interRAI-HC data [21,57]. In a previous study that also used machine learning to predict frailty status, logistic regression demonstrated comparable or higher performance in various scenarios [27]. This previous study suggested that the tree-based classifiers performed better if the data set was of low quality and contained bad features, and that MLP could generally show a greater performance if the data set is large enough and has complex structure with many layers. In our study, the reason why MLP did not show superior performance on average AUCs could be due to only 1 hidden layer being used.

On the other hand, when the number of features and sample sizes were large, machine learning models demonstrated better performance than logistic regression on both sensitivity and specificity. Specifically, MLP exhibited superior sensitivity, which means that it was more effective at accurately identifying frail older adults receiving home care and were at high risk of adverse health outcomes. In contrast, RF demonstrated superior specificity, which means that it was better at correctly identifying those who were not at high risk of adverse health outcomes. In the context of frailty, where interventions such as exercise, reduction of polypharmacy, and adequate nutrition can attenuate and even reverse the condition [52], high sensitivity is considered more important than high specificity if a trade-off between the 2 measures is required.

Our study revealed that the RF and XGBoost classifiers had significantly lower sensitivities and higher specificities than logistic regression, while MLP had higher sensitivities and lower specificities. This finding is consistent with previous studies on identifying frailty. For example, a study using various machine learning methods to develop predictive models for frailty conditions in older individuals based on an administrative health database [18] observed lower sensitivities and higher specificities for RF when predicting urgent hospitalization, and higher sensitivities and lower specificities for MLP when predicting various health outcomes, including mortality, fracture, and preventable hospitalization. Another similar study that developed a validated case definition of frailty using machine learning classifiers [27] found significantly lower sensitivities and higher specificities for XGBoost and RF compared to logistic regression on balanced data using the default threshold. These findings collectively suggest that identifying frailty using machine learning techniques remains challenging and future research is warranted to investigate the performance of machine learning models in other populations and care settings.

Implications for Research, Policy, and Practice

We did not identify any machine learning classifier that performed consistently better than the others. The best classifier differed across experimental conditions. Our results demonstrate that the advantages of using machine learning techniques to identify frail older adults become more apparent as the sample size and number of features increase. The logistic regression demonstrated higher or comparable AUC compared to machine learning classifiers in most scenarios. This differs from previous studies that show that machine learning classifiers outperformed logistic regression or its variants in predicting adverse health outcomes [18,20,24-26]. With a sample size of 95,042 and 138 features, Ridge logistic regression achieved an average AUC...
of 0.77 for 12-month mortality prediction. A logistic regression-based model developed by a previous study using interRAI-HC assessments of older persons in the New Zealand cohort targeting older individuals with complex comorbidities achieved an average AUC slightly higher (<0.01) than our result for 12-month mortality prediction [36]. The previous study used a slightly larger sample size of 104,436 and used a feature selection process to include only the features contributing over 1% to the performance. This may imply that a larger sample size and a feature selection process could further improve the predictive performance of logistic regression.

**Strengths and Limitations**

Our study used data collected from the interRAI instruments, standardized assessment instruments that have been developed by a collaborative network of health care professionals [21]. The interRAI instruments have been adopted in several jurisdictions to improve the quality of care for long-term care recipients, including Canada, Finland, Belgium, Italy, and Hong Kong. Therefore, the findings from this study may inform the identification of frail older adults for early interventions in similar care settings using interRAI assessments.

Our study has limitations. First, a successful measure of frailty should demonstrate satisfactory criterion validity against various adverse outcomes such as mortality, disability, hospitalization, and nursing home placement. Our study considered only mortality; therefore, it did not examine the accuracy of machine learning algorithms in predicting other adverse outcomes. Furthermore, we considered only 6- and 12-month mortality, resulting in an imbalanced data set that may yield higher specificity when using machine learning algorithms. It is also unclear whether the results can be extrapolated to other time intervals, such as 2 and 3 years. Further studies are needed to evaluate the prediction power of frailty against other critical outcomes. Second, the samples used in this study were limited to older adults receiving home care in New Zealand and most participants were Europeans. Future studies are warranted to assess the generalizability of this study’s findings. Third, we applied only 3 machine learning classifiers, chosen because they demonstrated better performance in several previous studies. The performance of other machine learning algorithms compared to regularized logistic regression was not investigated. Therefore, our conclusions are limited to the 3 algorithms examined. Fourth, calibration was not performed when training a machine learning classifier due to its additional computational costs, which may have affected the evaluation of model performance. The purpose of this study is to examine the impact of sample size and feature selection on the overall performance of each classifier in identifying frailty in older adults, rather than focusing on probability estimation or the quality of explanations provided by each model. It is worth noting that a recently published study [58] found that uncalibrated RF and XGBoost models performed similarly or even better than calibrated models in terms of accuracy and AUC. Therefore, the impact of calibration on our findings may not be severe. Finally, comparing the main features that affect the performance of different algorithms may improve the understanding of the construct of frailty. However, since the features in our simulation design were randomly selected across 1000 replications, the most important features identified from each run-in condition were not directly comparable. Therefore, we did not carry out further investigation on feature importance under different conditions.

**Conclusions**

Machine learning classifiers demonstrate considerable variability in prediction performance when assessed using different metrics. Regularized logistic regression is a reliable model for identifying frail older adults receiving home care, as indicated by the AUC, especially when the number of features and sample sizes are not excessively large. Conversely, MLP shows superior sensitivity, while RF demonstrates superior specificity when the number of features and sample sizes is large.

**Conflicts of Interest**

None declared.

Multimedia Appendix 1
Supplementary experiments: features and results.

[DOCX File, 2598 KB - ai_v3i1e44185_app1.docx ]

**References**


Abbreviations

- AUC: area under the curve
- CHS: Cardiovascular Health Study
- HC: home care
- interRAI-HC: interRAI-Home Care
- LASSO: Least Absolute Shrinkage and Selection Operator
- MLP: multilayer perceptron
- RF: random forest
- XGBoost: extreme gradient boosting

Edited by K El Emam, B Malin; submitted 09.11.22; peer-reviewed by C Bian, JR Medina, D Han; comments to author 02.07.23; revised version received 22.07.23; accepted 01.01.24; published 31.01.24.

Please cite as:
URL: https://ai.jmir.org/2024/1/e44185
doi:10.2196/44185
PMID:
Health Care Professionals’ and Parents’ Perspectives on the Use of AI for Pain Monitoring in the Neonatal Intensive Care Unit: Multisite Qualitative Study

Nicole Racine1*, PhD; Cheryl Chow2*, PhD; Lojain Hamwi2, BA; Oana Bucsea2, MA; Carol Cheng3, MSc; Hang Du4, MA; Lorenzo Fabrizi5, PhD; Sara Jasim5, MA; Lesley Johannsson6, BScN; Laura Jones5, PhD; Maria Pureza Laudiano-Dray7, MRes; Judith Meek7, MBBS, PhD; Neelum Mistry5, MSc; Vibhuti Shah8, MSc, MD; Ian Stedman9, LLB, PhD; Xiaogang Wang4, PhD; Rebecca Pillai Riddell2, PhD

1School of Psychology, University of Ottawa, Children’s Hospital of Eastern Ontario Research Institute, Ottawa, ON, Canada
2Department of Psychology, York University, Toronto, ON, Canada
3Department of Nursing, Mount Sinai Hospital, Toronto, ON, Canada
4Department of Mathematics and Statistics, York University, Toronto, ON, Canada
5Department of Neuroscience, Physiology and Pharmacology, University College London, London, United Kingdom
6Mount Sinai Hospital, Toronto, ON, Canada
7Neonatal Care Unit, University College London Hospitals, London, United Kingdom
8Department of Pediatrics, Mount Sinai Hospital, Toronto, ON, Canada
9School of Public Policy and Administration, York University, Toronto, ON, Canada
* these authors contributed equally

Corresponding Author:
Nicole Racine, PhD
School of Psychology, University of Ottawa
Children’s Hospital of Eastern Ontario Research Institute
136 Jean-Jacques Lussier
Ottawa, ON
Canada
Phone: 1 403 992 7869
Email: nracine2@uottawa.ca

Abstract

Background: The use of artificial intelligence (AI) for pain assessment has the potential to address historical challenges in infant pain assessment. There is a dearth of information on the perceived benefits and barriers to the implementation of AI for neonatal pain monitoring in the neonatal intensive care unit (NICU) from the perspective of health care professionals (HCPs) and parents. This qualitative analysis provides novel data obtained from 2 large tertiary care hospitals in Canada and the United Kingdom.

Objective: The aim of the study is to explore the perspectives of HCPs and parents regarding the use of AI for pain assessment in the NICU.

Methods: In total, 20 HCPs and 20 parents of preterm infants were recruited and consented to participate from February 2020 to October 2022 in interviews asking about AI use for pain assessment in the NICU, potential benefits of the technology, and potential barriers to use.

Results: The 40 participants included 20 HCPs (17 women and 3 men) with an average of 19.4 (SD 10.69) years of experience in the NICU and 20 parents (mean age 34.4, SD 5.42 years) of preterm infants who were on average 43 (SD 30.34) days old. Six themes from the perspective of HCPs were identified: regular use of technology in the NICU, concerns with regard to AI integration, the potential to improve patient care, requirements for implementation, AI as a tool for pain assessment, and ethical considerations. Seven parent themes included the potential for improved care, increased parental distress, support for parents regarding AI, the impact on parent engagement, the importance of human care, requirements for integration, and the desire for choice in its use. A consistent theme was the importance of AI as a tool to inform clinical decision-making and not replace it.
Conclusions: HCPs and parents expressed generally positive sentiments about the potential use of AI for pain assessment in the NICU, with HCPs highlighting important ethical considerations. This study identifies critical methodological and ethical perspectives from key stakeholders that should be noted by any team considering the creation and implementation of AI for pain monitoring in the NICU.

(JMIR AI 2024;3:e51535) doi:10.2196/51535

KEYWORDS
pain monitoring; pain management; preterm infant; neonate; pain; infant; infants; neonates; newborn; newborns; neonatal; baby; babies; pediatric; pediatrics; preterm; premature; assessment; intensive care; NICU; neonatal intensive care unit; HCP; health care professional; health care professionals; experience; experiences; attitude; attitudes; opinion; perception; perceptions; perspective; perspectives; acceptance; adoption; willingness; artificial intelligence; AI; digital health; health technology; health technologies; interview; interviews; parent; parents

Introduction

Globally, an estimated 13.4 million babies were born preterm in 2020, accounting for about 1 in 10 of all babies born [1]. Unfortunately, a significant proportion of preterm infants require neonatal intensive care unit (NICU) due to their vulnerability to complications and health issues [2]. As part of their lifesaving care, preterm infants undergo an average of 10 to 16 painful procedures per day [3]. Unmanaged NICU pain has significant developmental consequences [4,5] and is one of the largest sources of severe emotional distress in parents [6]. Pain assessment and management is a critical aspect of care in the NICU [7]. Traditional pain assessment methods in the NICU rely on observational tools [8,9]. However, there are several challenges with these methods, including bias and subjectivity, staff time resources, and potential variability in interpretation [10-12]. Given these challenges, innovative approaches are needed to improve existing pain assessment practices. Artificial intelligence (AI), which includes machine learning (ie, using a machine to extract knowledge from data and learn autonomously), is one technology that has shown tremendous potential in the health care field, and this potential may also inform the development of clinical decision support systems [13]. Specifically, AI-based technology can analyze large volumes of behavioral, physiological, and brain imaging data to provide suggestions with regard to infant pain assessment at the point of care.

Current evidence about the use of AI in the assessment and monitoring of infant pain appears to be promising [14,15]. Preliminary algorithms to monitor vital signs [16], such as heart rate, respiratory rate, and oxygen saturation, of preterm infants have been developed, all of which provide physiological indications of pain or distress as well as systems that incorporate behavioral indicators (eg, face movements, body movements, and crying) to predict pain [17]. Although there is immense potential for these new technologies to revolutionize how neonatal pain is assessed and monitored in the NICU, a limited understanding of the perspectives of key stakeholders with regard to this emerging technology exists, that is, health care professionals (HCPs) and parents. These perspectives are essential for the successful implementation of this technology in clinical practice.

Studies exploring the attitudes and trust of clinicians toward AI in health care found that while there is recognition of AI's potential benefits, concerns persist about reliability, transparency, data privacy, potential loss of autonomy in decision-making, and potential misinterpretation [18-21]. Factors such as age, education level, and previous experience with AI influenced attitudes and trust in AI technologies [21].

There is a growing interest in the application of AI technologies in health care, particularly in neonatal and pediatric care [14]. However, little is known about the perspectives of HCPs and parents on the use of AI for pain assessment in the NICU. Pain is a significantly different context warranting focused study because infants cannot verbalize for themselves. This study explores the perspectives of health care professionals and parents with regard to automated pain assessment using AI technology in the NICU. This study will inform the implementation of AI, specifically machine learning technology in the NICU, leading to more effective pain assessment and management strategies.

Methods

Ethical Considerations

Ethics approval for this qualitative study was granted from all study sites, including York University (2020-034), Mount Sinai Hospital (MSH; 19-0252-A), and University College London Hospital (UCLH; 11/LO/0350). Informed consent was obtained from all participants. All data were deidentified. Individuals were provided with a CAD $10 (approximately US $7) gift card to a local coffee shop for their participation.

Setting and Design

Data collection occurred at 2 tertiary care NICUs: MSH (Toronto, Canada) and UCLH (London, United Kingdom). The study is part of a larger project focused on the use of AI, specifically the development of a machine learning algorithm, to assess infant pain in the NICU. Participants consisted of 20 HCPs (nurses, physicians, and allied health professionals) and 20 parents (mothers and fathers). Recruitment at MSH took place from February to March 2020, and recruitment at UCLH took place from July 2021 to October 2022. Interviews at MSH occurred in person at the hospital, whereas interviews at UCLH were web-based and conducted using a secure Zoom platform (Zoom Video Communications). This difference was due to the onset of the COVID-19 pandemic after the study had launched, which delayed the UK interviews and necessitated the use of a secure web platform. For HCPs, eligibility criteria were (1) currently providing care to infants at one of the NICUs and (2)
trained as either a nurse, physician, or other health professionals (ie, outreach staff and consultant practice educator). For parents, eligibility criteria included being 18 years and older of age, having an infant who was currently receiving care in the NICU, and being fluent in English, orally (in order to respond to complex questions in the interview). Using a purposive sampling approach, all participants were initially approached by 1 clinical member of staff on the unit and asked if they were interested in participating in the study. Only families where the parent was at least 18 years of age and spoke English were approached. If interested, they received additional information, and a time was scheduled for an interview.

Following introductions and the completion of the consent form, 30-minute semistructured interviews were conducted by a member of the research team (NR, C Chow, and L Johannsson) in a private clinic room (MSH) or web-based room (UCLH). Baseline demographic information was collected at the outset of the meeting followed by a series of questions (10 for HCPs and 9 for parents) pertaining to the use of AI to inform NICU decision-making related to the assessment of infant pain. Notes were taken during the interviews to supplement transcripts. Interviewers read an initial script providing a definition of AI and providing context for the study. In-person interviews were recorded using a digital audio recorder, whereas web-based interviews were recorded using privacy-compliant web software (Zoom) and stored on a secure server. All participants were debriefed following the interview and provided with a gift card to a local coffee shop as a token of appreciation. Standards for Reporting Qualitative Research were followed for this study (Multimedia Appendix 1 [22]).

Development of the Interview Guides

Using a grounded theory approach [23], the goal of the qualitative interviews was to generate detailed knowledge about HCPs’ and parents’ understandings and perceptions of the use of AI in the NICU to assist with infant pain assessment and management. Specifically, we sought to gain insight into HCPs’ and parents’ understanding of AI, perceived implications of this technology, potential benefits of the technology, and barriers to its use in the NICU setting. Two interview guides were developed to address the diverse perspectives of HCPs (Multimedia Appendix 2) and parents (Multimedia Appendix 3). The interview guides were developed collaboratively by members of the research team (RPR and NR), who are clinical psychologists with previous experience in conducting qualitative research with both HCPs and parents in the NICU and other pediatric medical settings [24,25]. The guides were reviewed and edited based on the feedback from team members with NICU clinical expertise (VS, C Chow, JM, and MPL-D) as well as ethical or legal or social expertise related to AI (IS). Interviews were conducted by 2 postdoctoral fellows (NR and C Chow) and 1 research staff (L Johannsson). A decision was made in advance to review and make necessary changes to the questions after the first interviews were conducted at each site based on participant comprehension and feedback. Based on the review, no major alterations were required. Participants had the opportunity to provide any additional comments or feedback at the end of the interview. Interviews were conducted until saturation was reached [26].

Data Processing and Analysis

The interview audio recordings were anonymized and transcribed by 1 research assistant and independently double-checked by members of the research team. Transcripts were subsequently analyzed using 6 phases of thematic analysis (ie, familiarization, generating codes, identifying themes, reviewing themes, naming themes, and report writing) [27]. Data analyses took place from February to April 2023. There were 3 analysis leads (NR, C Chow, and RPR) who took primary responsibility for developing the code book, overseeing the coding process, and developing themes based on the codes generated. As a first step, the analysis leads familiarized themselves with the data by reading and making notes on the transcripts. Responses were examined for differences between the 2 sites (eg, unique considerations related to the country, time, or modality via in-person vs web-based) or any effects that may have necessitated a different analysis pathway. It was determined that there were no differences, and we proceeded with analyzing the transcripts together. Next, a list of initial codes was generated independently by the analysis leads prior to a consensus meeting. Two consensus meetings were held, where all codes were reviewed and agreed upon. Subsequently, the analysis leads (NR, RPR, and C Chow) ran a 90-minute training session with 10 coders to familiarize them with the codes that have been created. All coders (LH, SJ, OB, VS, MPL-D, C Cheng, IS, HD, NM, and L Jones) were members of an interdisciplinary research team (ie, neurobiology, behavioral neuroscience, neurophysiology, psychology, medicine, nursing, and law) with research backgrounds in pediatric health care, with most specializing in infant care. Each transcript was coded twice. The average percent agreement (ie, the number of times 2 individuals agreed upon a code divided by the total number of units of observation that were rated) across transcripts between coders for the HCP and parent transcripts was 0.77, which is adequate [28]. Next, the analysis leads reviewed the coded transcripts and collated codes for each question. The analysis leads met and generated relevant potential themes and a thematic map based on the data. Finally, examples were selected to accompany each theme, which are presented in the results below. Summary statistics of all demographic variables were conducted in SPSS (version 28; IBM Corp).

Results

Participant Characteristics

The participant characteristics are shown in Table 1. In total, 90% (n=18) of HCPs were university-educated and had extensive experience in the NICU (mean 19.4, SD 10.69 years; range 4-37 years). For HCPs, 55% (n=11) reported “Western” cultural heritages (eg, Canadian, British, and Australian), 5% (n=1) African, 15% (n=3) East Asian, 10% (n=2) Caribbean, 10% (n=2) South Asian, and 5% (n=1) not reported. For parents, 80% (n=16) reported “Western” cultural heritages (eg, Canadian, European, or Australian), 5% (n=1) Asian, 5% (n=1) Middle Eastern, and 10% (n=2) not reported. Most parents who participated across both sites were mothers (n=17, 85%) with a mean age of 34 (SD 5.42) years. In total, 90% (n=18) of parents had a university education or higher.
Table 1. Participant demographic characteristics.

<table>
<thead>
<tr>
<th>Characteristics</th>
<th>Health care providers (n=10 each)</th>
<th>Parents (n=10 each)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Mount Sinai Hospital</td>
<td>University College London Hospital</td>
</tr>
<tr>
<td>Gender, n (%)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Women</td>
<td>9 (90)</td>
<td>8 (80)</td>
</tr>
<tr>
<td>Men</td>
<td>1 (10)</td>
<td>2 (20)</td>
</tr>
<tr>
<td>Age (years), mean (SD)</td>
<td>_a</td>
<td>—</td>
</tr>
<tr>
<td>Postnatal age of infant (days), mean (SD)</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td>Highest level of education, n (%)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Graduate school or professional training</td>
<td>6 (60)</td>
<td>7 (70)</td>
</tr>
<tr>
<td>University graduate</td>
<td>2 (20)</td>
<td>3 (30)</td>
</tr>
<tr>
<td>Partial university</td>
<td>0 (0)</td>
<td>0 (0)</td>
</tr>
<tr>
<td>Trade school or community college</td>
<td>1 (10)</td>
<td>0 (0)</td>
</tr>
<tr>
<td>High school graduate</td>
<td>0 (0)</td>
<td>0 (0)</td>
</tr>
<tr>
<td>Less than high school</td>
<td>0 (0)</td>
<td>0 (0)</td>
</tr>
<tr>
<td>Not reported</td>
<td>1 (10)</td>
<td>0 (0)</td>
</tr>
<tr>
<td>Heritage culture, n (%)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>African</td>
<td>1 (10)</td>
<td>0 (0)</td>
</tr>
<tr>
<td>Asian</td>
<td>2 (20)</td>
<td>1 (10)</td>
</tr>
<tr>
<td>Australia or New Zealand</td>
<td>0 (0)</td>
<td>0 (0)</td>
</tr>
<tr>
<td>Caribbean</td>
<td>2 (20)</td>
<td>0 (0)</td>
</tr>
<tr>
<td>Canadian</td>
<td>1 (10)</td>
<td>0 (0)</td>
</tr>
<tr>
<td>European</td>
<td>3 (30)</td>
<td>7 (70)</td>
</tr>
<tr>
<td>Middle Eastern</td>
<td>0 (0)</td>
<td>0 (0)</td>
</tr>
<tr>
<td>South Asian</td>
<td>0 (0)</td>
<td>2 (20)</td>
</tr>
<tr>
<td>Not reported</td>
<td>1 (10)</td>
<td>0 (0)</td>
</tr>
<tr>
<td>Type of health care professional, n (%)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Physician</td>
<td>5 (50)</td>
<td>3 (30)</td>
</tr>
<tr>
<td>Registered nurse</td>
<td>5 (50)</td>
<td>4 (40)</td>
</tr>
<tr>
<td>Other health professional</td>
<td>0 (0)</td>
<td>3 (30)</td>
</tr>
<tr>
<td>Experience (years), mean (SD)</td>
<td>22 (8.55)</td>
<td>16 (12.18)</td>
</tr>
</tbody>
</table>

*Not available.

**HCP Themes**

Six themes emerged from the thematic analysis on the HCP interviews. Each theme, a description, and representative quotes are presented in Table 2. HCP themes and subthemes are presented in Figure 1. First, in the context of their comfort with incorporating new AI technology, HCPs reported limited experience with AI technology in the NICU (1 HCP was part of a research study at another institution), and they were comfortable using other forms of technology. Second, HCPs identified some concerns with regard to the integration of AI for pain assessment in the NICU. Some of these concerns included increased distress from knowing clinicians were inflicting pain and extra workload for HCPs, increased stress for parents, and decreased opportunities for parent-child bonding, as well as fears related to overreliance on AI technology and the overuse of medication to manage pain. Despite these concerns, the third theme emerged surrounding several benefits that AI could bring to the NICU context. Notably, HCPs identified increased awareness of infant pain, early detection and diagnosis of clinical changes, increased efficiency, and standardization of pain assessment, as well as the potential to inform the development of better pain management strategies. From a practical standpoint, the fourth theme identified requirements to facilitate the implementation of AI in the NICU, including the size of machinery, staff training, as well as clearly communicating the validity, sensitivity, and specificity of the algorithm being used. The fifth...
The theme that was unanimously shared was the idea that using AI for pain assessment in the NICU would be a tool for HCPs to use but could not replace the clinical judgment and decision-making of an HCP. Concerns related to how the next generation of HCPs would be trained to ensure that they have both the clinical and technological skills to operate in the NICU were described, given the potential overreliance on technology. Finally, HCPs identified the potential for ethical concerns related to an AI algorithm for constant pain monitoring in the NICU, specifically, issues related to the disagreement between HCP and the AI algorithm, implications of pain monitoring in the absence of pain management, as well as the need to audit the algorithm. Overall, there was general acceptability for the benefits, use, and integration of AI technology for pain assessment in the NICU, with keen identification of the potential work-related, structural, technological, and ethical issues that would need to be addressed to facilitate implementation.

Figure 1. Themes and subthemes generated from qualitative interviews with HCPs on their perspectives about using AI to assess pain in the NICU. AI: artificial intelligence; Ax: assessment; HCP: health care professional; NICU: neonatal intensive care unit.
Table 2. Key themes identified by HCPs\(^a\) with regard to the use and integration of AI\(^b\) for pain assessment in the NICU\(^c\).

<table>
<thead>
<tr>
<th>Theme</th>
<th>Description</th>
<th>Representative quote</th>
</tr>
</thead>
<tbody>
<tr>
<td>Technology is used regularly in the NICU</td>
<td>HCPs shared that despite having limited experience with AI specifically, they use technology to inform their clinical decision-making and they feel comfortable using the technology that is currently available.</td>
<td>• “It informs everything. I think that’s one of the things that working in intensive care is that we use technology and monitoring to inform a lot of our decisions.”</td>
</tr>
<tr>
<td>Concerns of AI integration for pain assessment in the NICU</td>
<td>HCPs identified concerns related to the integration of AI in the NICU. It specifically increased the workload for HCPs and increased distress, knowing they were potentially inflicting pain on an infant. They also reported that constant pain monitoring could increase stress for parents and that additional machinery could inhibit parent-child bonding. Concerns were also identified with regard to the over-reliance on what the algorithm reported and the overuse of pain pharmaceuticals to manage pain.</td>
<td>• Increased HCP distress: “I’m not sure cause you imagine like how upsetting it would be like you know I’m doing a diaper change and this thing is telling me the baby is in pain.” • Increased workload: “I think there would be some negative feedback towards having extra work to be done.” • Fear of over-reliance on the AI: “The disadvantages would be that we become over reliant on it. And just because the machine says the baby’s not in pain, then it could be dismissed as the baby isn’t in pain, when actually if you look at the baby, you can tell they’re in pain.” • Increased parent stress: “It can cause stress ... Unnecessary stress.” • Impeding parent-child bonding: “I can see it taking away from looking at babies…you see parents, particularly looking at their monitor alarms, for whatever reason, they look more at the monitor than actually what their baby’s doing.”</td>
</tr>
<tr>
<td>AI has the potential to improve pain assessment and management</td>
<td>HCPs indicated there are several ways in which integrating constant pain monitoring in the NICU could improve clinical care, including the development of new therapies, early diagnosis of difficulties, detection of changes in clinical presentation, increased awareness of infant pain, increased efficiency of pain assessment, increased standardization of pain assessment, and increased collaboration between HCPs and parents.</td>
<td>• “I think it’s good that um there is a form of technology that can give us more information about pain in this population because I think there’s a lot of unknown and I think well I know for myself like I said I can’t honestly say that I’m always thinking about if this baby is in pain or what kind of pain this baby is in when doing a procedure.” • “I think it would give them more time to obviously focus on other aspects of their work instead of having to score every half an hour or so to proceed and enter the data as it is at the moment.”</td>
</tr>
<tr>
<td>Requirements for implementation of AI in NICU</td>
<td>HCPs described structural (ie, machine size and invasiveness of machinery) requirements for implementing AI in the NICU. Specifically, machinery would need to be small and non-invasive. HCPs indicated that training staff to understand and interpret the output provided by the technology is important. They also indicated that the algorithm would need to be properly validated and sensitive for detecting pain in diverse patient groups and situations.</td>
<td>• Structural requirements: “It depends how invasive the technology is. When you have a 450 gram baby in front of you. Even putting on things like more monitors actually occludes your that visual assessment of the child. So I think there can be barriers.” • Importance of training: “I think obviously it’s all about training ... everybody understands how it works and the benefits.”</td>
</tr>
<tr>
<td>AI is a tool to inform clinical pain assessment and management</td>
<td>HCPs indicated that AI in the NICU should be viewed as a tool to inform clinical decision-making but not as a replacement. They also indicated that the integration of this technology would have implications for the training of new HCPs to ensure they have the ability to understand how this tool could inform their own clinical assessment.</td>
<td>• “I like using technology but as long as it doesn’t replace my ability to provide comfort and care” • “If I’m gonna make it’s just detection of pain, I think it’d be fairly comfortable with that. Because then I can react to that. Whereas if it’s making medical decision on the treatment, a baby’s receiving, I think that will be a completely different scenario.”</td>
</tr>
<tr>
<td>Ethical concerns with constant pain monitoring may occur</td>
<td>HCP indicated the need to be aware of ethical concerns like the potential bias in AI algorithms, disagreements between HCPs and the AI’s output, and the implications of constant pain monitoring without intervening. HCPs also indicated that algorithms would need to be audited and monitored over time.</td>
<td>• “And then you have to decide, what you want to do about it. And then you have to decide, in a medical-legal issue whether to believe A.I. or the clinician and that will be interesting.”</td>
</tr>
</tbody>
</table>

\(^a\)HCP: health care professional.  
\(^b\)AI: artificial intelligence.  
\(^c\)NICU: neonatal intensive care unit.
Parent Themes

Seven overarching themes were identified with parents (Table 3). Parent themes and subthemes are presented in Figure 2. First, parents indicated it would be desirable to know if their infants were in pain because there are limited ways of assessing neonatal pain and it would provide useful information to HCPs to improve their infant’s care. However, the second theme arose about the emotional toll that may be experienced by parents. Some parents noted heightened distress from knowing their infant was experiencing pain. The third theme revolved around a preference to have parents decide for themselves whether they wanted continuous pain monitoring using AI. The fourth theme was that parents indicated wanting support to interpret and understand the constant pain monitoring. That is, they would want HCPs to explain their decision-making process as well as how the pain assessment provided by the AI was being used.

The fifth theme was that parents perceived their current level of engagement in their infant’s care to be quite high and they did not think constant pain monitoring would change this engagement. The sixth theme was that most parents would not trust an AI to make an independent decision about their infant’s pain but rather believe it should be incorporated as a tool by HCPs to make a clinical decision. Parents voiced that there would be potential for error in the AI’s assessment and that verification by an HCP would be important. Finally, parents identified requirements related to AI integration in the NICU. Specifically, they are concerned about privacy since large amounts of data would be collected and therefore would need to be kept secure. They also identified that the algorithm should be developed in a nonbiased way and that generalizability of the algorithm across infant presentations and contexts would be needed.

Figure 2. Themes and subthemes generated from qualitative interviews with parents on their perspectives of using AI to assess pain in the neonatal intensive care unit. AI: artificial intelligence; HCP: health care professional.
<table>
<thead>
<tr>
<th>Theme</th>
<th>Description</th>
<th>Representative quote</th>
</tr>
</thead>
<tbody>
<tr>
<td>Constant pain monitoring can facilitate better care</td>
<td>Parents indicated there are advantages to constant pain monitoring (eg, increase in awareness of infant’s experience and confidence in care provided).</td>
<td>“But then it could also help the parent, could help us understand the baby a bit more and maybe bond maybe a bit more or communicate in a way with the baby more.”</td>
</tr>
<tr>
<td>Emotional distress may result from constant pain monitoring</td>
<td>Parents shared disadvantages to constant pain monitoring, such as too much information or distress associated with knowing their child is in pain.</td>
<td>“My gut is saying, as a parent, well, of course. But I’m wondering whether you can have almost too much information, where if certain things, I definitely would be in this position, where if certain things had to be done to my child, life and death or even just less serious, but they needed to be done for you, know, health reasons, how productive is it for a parent to know exactly how much pain their child is in.”</td>
</tr>
<tr>
<td>Desire for choice in using the constant pain monitoring</td>
<td>Parents indicated that they would like to be given a choice to view the constant pain monitoring.</td>
<td>“You should have a choice in the same way as like, you can choose to look at lots of the information about your baby or not.”</td>
</tr>
<tr>
<td>Desire for support related to AI</td>
<td>Parents indicated that they would want communication from staff and support to understand and interpret the constant pain monitoring. They would also like basic information about how the algorithm was developed and makes its predictions.</td>
<td>“Because even now I don’t want to do anything unless the nurse is there ... but you see that number go up as you’re as you’re caring for the baby you might be or I might be a little apprehensive um but with the reassurance of the nurse or if you can see that once the baby is settled down the baby is more comfortable again then you know that it’s ok.”</td>
</tr>
<tr>
<td>Minimal impact on parent engagement</td>
<td>Parents indicated that constant pain monitoring would minimally impact their level of active engagement in the newborn’s care as most reported that they already engaged at a high level.</td>
<td>“I would want to know, and I would want it to be very clear why those decisions were made. I would want, if we were using kind of artificial intelligence, what kind of almost a report on why those decisions were made and why it was recommended that XYZ happened as a result.”</td>
</tr>
<tr>
<td>AI for pain monitoring is a tool for HCPs[^1], not a replacement</td>
<td>Parents indicated that constant pain monitoring should be used as a tool to inform clinical judgment.</td>
<td>“You know I’m not sure that it would change how engaged I would be because I think you know you can use other metrics as like surrogate of pain as well and being at the bedside you can still be engaged in her care but I guess it could be interested to ask you know like when we should up like you know how were her pain scores overnight or something like that. And you know get that data and get that information from the bedside nurse. But I don’t think it would dramatically change the engagement.”</td>
</tr>
</tbody>
</table>

[^1]: JMIR AI 2024 | vol. 3 | e51535 | p.104 https://ai.jmir.org/2024/1/e51535 (page number not for citation purposes)
Discussion

Principal Findings

This international study includes the perspectives of both HCPs (ie, physicians and nurses) and parents regarding the use of AI technology in the NICU setting. These perspectives offer critical insights to help inform the development of potential AI technology on infant pain management and integration of this technology as part of clinical decision support systems. We found that both HCPs and parents were supportive of the use of AI technology in predicting infant pain. Both HCPs and parents recognized that AI has the potential to improve care in the NICU setting. Other studies have also identified similar benefits including earlier detection of illness, increased collaboration and communication, and development of new treatments that further support the use of AI in clinical settings [29,30].

In line with previous research [31], this study also found that HCPs and parents had similar concerns on the use of AI technologies in the NICU setting, including effectiveness and accuracy, fear of overreliance, and shared decision-making over the use of AI technology. Furthermore, we identified additional themes from the perspectives of parents regarding the importance of receiving support for interpreting and understanding constant pain monitoring. Interestingly, most parents indicated that they would prefer the choice to have access to constant pain monitoring in real time, as it could impact parents differently. Moreover, both HCPs and parents identified the importance of using AI as an adjunctive tool to inform clinical decisions. That is, both parents and HCPs seemed in favor of using AI to augment human intelligence and support more informed clinical decision-making [32] rather than automating any aspect of clinical care. Similar to youth and adult patients, parents of infants in the NICU were concerned about the risk of clinician replacements and emphasized the importance of the human element (ie, HCP’s presence at the bedside) in clinical care [30,33,34]. Clinicians also warned about the potential for diminished skills and overreliance on technology for the next generation of clinicians with regard to pain assessment at the bedside. It is worth noting that clinical decision-making and responsibility continue to rest with clinicians, and there is currently no legislation that would allow automated health care decisions by an AI [35]. These new emerging themes could potentially help inform the future development of AI tools in the NICU setting as well as the training of future HCPs working in the NICU. Findings from this study could be used to justify increased training, engagement, and consultation with health care professionals as AI is implemented in the NICU.

Interestingly, we found very similar responses and results across countries as well as interview modalities. This is not surprising as both the United Kingdom and Canada follow similar protocols within the NICUs as both have public health care systems. Additionally, structured interviews, such as those conducted in this study, work equally well in face-to-face or web-based studies [36]. Furthermore, the interviewers were the same across both contexts. We also found that both HCPs and parents had limited experience with the use of AI in the NICU, meaning that all the responses garnered in this study were hypothetical in nature. Had participants had exposure, they may have provided different responses with regard to the feasibility and use of this technology. Future research prior to and during the implementation process will be important to capture these perspectives.

Limitations

There are some limitations to this study that should be considered when interpreting our results. First, interviews were conducted with HCPs at 2 large, tertiary-care, academic hospitals in Canada and the United Kingdom that are at the forefront of technological advancement in the NICU. As such, the perspectives of HCPs in this study may not be generalizable to smaller, less well-resourced care settings. Second, parents included in this study were highly educated, which may limit generalizability to parents with lower educational attainment, which is also a known risk factor for preterm birth [37]. Moreover, parents were recruited into the sample if they spoke English, which may have resulted in a less culturally diverse sample. Third, many of the themes that were identified by HCPs

Requirements for AI integration in the NICU

Parents indicated that it would be important to consider how data might be collected and used by the AI, how to reduce bias in the development of the algorithm, and how to ensure that the algorithm was generalizable across infants and contexts.

- “... questions about the data it was collecting and where that was going and who’s using that data. So obviously, the monitoring there’s a lot of information there.”
- “I would be concerned if a model was created that the way in which it was created was maybe not ethical but I’m know there’s all kinds of laws and things like that but I was just thinking about how that might work.”
- “And then the sample size and the how many different like every baby is different and every baby’s pain tolerance is different how do you know that you’ve got all your bases covered for all the different scenarios.”

<table>
<thead>
<tr>
<th>Theme</th>
<th>Description</th>
<th>Representative quote</th>
</tr>
</thead>
</table>
| Requirements for AI integration in the NICU | Parents indicated that it would be important to consider how data might be collected and used by the AI, how to reduce bias in the development of the algorithm, and how to ensure that the algorithm was generalizable across infants and contexts. | - “… questions about the data it was collecting and where that was going and who’s using that data. So obviously, the monitoring there’s a lot of information there.”
- “I would be concerned if a model was created that the way in which it was created was maybe not ethical but I’m know there’s all kinds of laws and things like that but I was just thinking about how that might work.”
- “And then the sample size and the how many different like every baby is different and every baby’s pain tolerance is different how do you know that you’ve got all your bases covered for all the different scenarios.” |

aAI: artificial intelligence.
bNICU: neonatal intensive care unit.
cHCP: health care professional.
and caregivers were broad in that they were not referring to the use of AI specifically but rather the use of clinical decision support systems (ie, a clinician using technology like AI to help inform their decisions related to care). As both technology and terminology evolve in the medical context, it will be important to disentangle opinions related to the technology itself as opposed to its use as a clinical decision-making tool. Finally, questions asked of HCPs and parents differed with more emphasis placed on general technology with HCPs and on neonatal pain for parents. This may have had an impact on the responses that were generated. As AI-related technology is integrated into medical settings, future qualitative research may focus specifically on pain-related questions.

Conclusions

Based on detailed interviews with 40 HCPs and parents across 2 large NICUs in publicly funded hospitals in Canada and the United Kingdom, our overall findings indicate that both HCPs and parents view the integration of an AI algorithm for constant pain monitoring to have potential benefits and to be an acceptable practice. Notably, HCPs identified several ways in which constant pain monitoring could improve the clinical care provided in the NICU. Both HCPs and parents were balanced in their perspectives and identified potential disadvantages as well as requirements for the successful implementation of an AI tool for pain assessment. Taken together, there is immense promise as well as major structural, ethical, and methodological considerations for the development and implementation of AI technology in the NICU setting.

Acknowledgments

This project was funded by the Canadian Institutes of Health Research, Natural Sciences and Engineering Research Council, Social Sciences and Humanities Research Council, and Collaborative Health Research Projects (principal investigator: RPR; grant 1001-2019-0004). LF, JM, L Jones, and MPL-D were funded by the Medical Research Council UK (grant MR/S003207/1). The funders had no role in data collection, interpretation, and reporting. NR receives funding through the Chair in Child and Youth Mental Health at the Children’s Hospital of Eastern Ontario and the University of Ottawa.

Authors’ Contributions

RPR, NR, and C Chow conceptualized the paper, developed the qualitative coding system, conducted the qualitative interviews, did the data analysis, and wrote the first draft of the paper. LH contributed to writing the introduction. L Johannsson and C Cheng contributed to data collection. LH, OB, C Cheng, HD, LF, SJ, L Johannsson, L Jones, MPL-D, JM, NM, VS, IS, and XW contributed to data coding. All authors reviewed the final paper. RPR was responsible for obtaining the primary funding.

Conflicts of Interest

None declared.

Multimedia Appendix 1

Standards of Reporting Qualitative Research [22].

[DOCX File, 19 KB - ai_v3i1e51535_app1.docx ]

Multimedia Appendix 2

Interview guide for health care professionals.

[DOCX File, 16 KB - ai_v3i1e51535_app2.docx ]

Multimedia Appendix 3

Interview guide for caregivers.

[DOCX File, 16 KB - ai_v3i1e51535_app3.docx ]

References


Abbreviations

- AI: artificial intelligence
- HCP: health care professional
- MSH: Mount Sinai Hospital
- NICU: neonatal intensive care unit
- UCLH: University College London Hospital

©Nicole Racine, Cheryl Chow, Lojain Hamwi, Oana Bucsea, Carol Cheng, Hang Du, Lorenzo Fabrizi, Sara Jasim, Lesley Johannsson, Laura Jones, Maria Pureza Laudiano-Dray, Judith Meek, Neelum Mistry, Vibhuti Shah, Ian Stedman, Xiaogang Wang, Rebecca Pillai Riddell. Originally published in JMIR AI (https://ai.jmir.org), 09.02.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR AI, is properly cited. The complete bibliographic information, a link to the original publication on https://www.ai.jmir.org/, as well as this copyright and license information must be included.
Reidentification of Participants in Shared Clinical Data Sets: Experimental Study

Daniela Wiepert1, BA; Bradley A Malin2,3,4, PhD; Joseph R Duffy1, PhD; Rene L Utianski1, PhD; John L Stricker1, PhD; David T Jones1, MD; Hugo Botha1, MBChB

1Department of Neurology, Mayo Clinic, Rochester, MN, United States
2Department of Biomedical Informatics, Vanderbilt University Medical Center, Nashville, TN, United States
3Department of Biostatistics, Vanderbilt University Medical Center, Nashville, TN, United States
4Department of Computer Science, Vanderbilt University Medical Center, Nashville, TN, United States

Corresponding Author:
Hugo Botha, MBChB
Department of Neurology
Mayo Clinic
200 1st St
Rochester, MN, 55905
United States
Phone: 1 5072841588
Email: botha.hugo@mayo.edu

Abstract

Background: Large curated data sets are required to leverage speech-based tools in health care. These are costly to produce, resulting in increased interest in data sharing. As speech can potentially identify speakers (ie, voiceprints), sharing recordings raises privacy concerns. This is especially relevant when working with patient data protected under the Health Insurance Portability and Accountability Act.

Objective: We aimed to determine the reidentification risk for speech recordings, without reference to demographics or metadata, in clinical data sets considering both the size of the search space (ie, the number of comparisons that must be considered when reidentifying) and the nature of the speech recording (ie, the type of speech task).

Methods: Using a state-of-the-art speaker identification model, we modeled an adversarial attack scenario in which an adversary uses a large data set of identified speech (hereafter, the known set) to reidentify as many unknown speakers in a shared data set (hereafter, the unknown set) as possible. We first considered the effect of search space size by attempting reidentification with various sizes of known and unknown sets using VoxCeleb, a data set with recordings of natural, connected speech from >7000 healthy speakers. We then repeated these tests with different types of recordings in each set to examine whether the nature of a speech recording influences reidentification risk. For these tests, we used our clinical data set composed of recordings of elicited speech tasks from 941 speakers.

Results: We found that the risk was inversely related to the number of comparisons an adversary must consider (ie, the search space), with a positive linear correlation between the number of false acceptances (FAs) and the number of comparisons ($r=0.69$; $P<.001$). The true acceptances (TAs) stayed relatively stable, and the ratio between FAs and TAs rose from 0.02 at $1 \times 10^5$ comparisons to 1.41 at $6 \times 10^6$ comparisons, with a near 1:1 ratio at the midpoint of $3 \times 10^6$ comparisons. In effect, risk was high for a small search space but dropped as the search space grew. We also found that the nature of a speech recording influenced reidentification risk, with nonconnected speech (eg, vowel prolongation: FA/TA=98.5; alternating motion rate: FA/TA=8) being harder to identify than connected speech (eg, sentence repetition: FA/TA=0.54) in cross-task conditions. The inverse was mostly true in within-task conditions, with the FA/TA ratio for vowel prolongation and alternating motion rate dropping to 0.39 and 1.17, respectively.

Conclusions: Our findings suggest that speaker identification models can be used to reidentify participants in specific circumstances, but in practice, the reidentification risk appears small. The variation in risk due to search space size and type of speech task provides actionable recommendations to further increase participant privacy and considerations for policy regarding public release of speech recordings.
KEYWORDS
reidentification; privacy; adversarial attack; health care; speech disorders; voiceprint

Introduction

Background

Advances in machine learning and acoustic signal processing, along with widely available analysis software and computational resources, have resulted in an increase in voice- and speech-based (hereafter referred to as speech for simplicity) diagnostic and prognostic tools in health care [1]. Applications of such technology range from the early detection of cardiovascular [2], respiratory [3], and neurological [4] diseases to the prediction of disease severity [5] and evaluation of response to treatment [6]. These advances have substantial potential to enhance patient care within neurology given the global burden of neurological diseases [7,8], the poor global access to neurological expertise [9,10], and the established role of speech examination within the fields of neurology and speech-language pathology [11].

Large curated data sets are needed to harness the advances in this area. These data sets are costly to assemble and require rare domain expertise to annotate, leading to increased interest in data sharing among investigators and industry partners. However, given the potentially identifiable nature of voice or speech recordings and the health information contained within such recordings, significant privacy concerns emerge. For many data sets, conventional deidentification approaches that remove identifying metadata (eg, participant demographics and date and location of recording) are sufficient, but sharing speech recordings comes with additional risk as the speech signal itself has the potential to act as a personal identifier [12-14]. In recognition of this potential problem, voiceprints are specifically mentioned as an example of biometric identifiers with respect to the Health Insurance Portability and Accountability Act (HIPAA) Privacy Rule [15,16]. Approaches that involve modifying nonlinguistic aspects of speech through distortion or alteration of the signal may address the inherent identifiability of the speech signal (ie, its potential as a voiceprint) [13,17], but this is not an option when a central part of speech examination in medicine is to use the acoustic signal to detect subtle nonlinguistic abnormalities indicative of the presence of neurological disease [11,13]. Deidentification in compliance with HIPAA may still be possible under the Expert Determination implementation, whereby the risk of reidentification for unmodified speech recordings is deemed low according to accepted statistical and scientific principles [15,16]. In this respect, various previous studies have investigated the risk of reidentification in research cohort data sets based on demographic or other metadata that may link a participant to their corresponding recordings [18-20], but none have explicitly assessed the inherent risk of the acoustic signal itself. Determining the risk of reidentification for recordings in speech data sets and learning how to best mitigate such risk is necessary for health care institutions to protect patients, research participants, and themselves.

Unfortunately, the same machine learning advances that facilitate the use of speech in health care have also made adversarial attacks, such as deanonymization or reidentification attacks, more feasible. For example, attempting to reidentify a speaker from only a speech recording relies on the mature, well-researched field of speaker identification [21,22]. Studies using speaker identification suggest that the potential for identification from the acoustic signal alone is high [23], although there have been minimal studies in the context of adversarial attacks that may result in potential harm to a speaker [24,25]. Only one previous study has relied on a speaker identification model for reidentification, and the results suggested that the risk was high with a single unknown or unidentified speaker and a moderately small reference set of 250 known or identified speakers [25]. As such, the risk inherent in the acoustic signal, devoid of metadata, is nonzero but relatively unknown, and the feasibility for larger data sets is unexplored.

In addition, these approaches are rarely applied to medical speech data sets [26]. This presents a gap in research as medical speech recordings differ from speech recordings of healthy speakers in a few systematic ways. First, the recordings typically contain speech with abnormalities (ie, speech disorders), which may make reidentification harder as many speech disorders are the result of progressive neurological disease, which causes changes in speech that evolve over months to years [11]. Matching recordings from a time when a speaker was healthy or mildly affected to recordings in which they have a more severe speech disorder may be more difficult [27-29]. Second, the premise of speaker identification is that there are recognizable between-speaker differences tied to identity. However, in a cohort enriched with speech with abnormalities, a substantial proportion of the variance would be tied to the underlying speech disorder as this causes recognizable deviations [11], resulting in speakers sounding less distinct [30]. Finally, medical speech recordings typically contain responses to elicited speech tasks rather than the unstructured connected speech typically used in identification experiments. Some speech task responses do contain connected speech (eg, paragraph reading), but others are very dissimilar (eg, vowel prolongation). The impact of speech task on identifiability remains unknown.

Objectives

In this study, we addressed the risk of reidentification in a series of experiments exploring the reidentifiability of medical speech recordings without using any metadata. We accomplished this goal by modeling an adversarial attack using a state-of-the-art speaker identification architecture wherein an adversary trains the speaker identification model on publicly available, identified recordings and applies the model to a set of unidentified clinical recordings.

https://ai.jmir.org/2024/1/e52054
**Methods**

**Overview**

Our experimental design was based on the following assumptions: (1) a data recipient has decided to attempt reidentification of study participant data, thereby becoming an adversary; and (2) this adversary relies on an adversarial attack strategy known as a marketer attack, wherein they use a large data set of identified speech (hereafter referred to as the known set), perhaps obtained from a web source such as YouTube, to train a speaker identification model that is then used to reidentify as many unknown speakers in the shared clinical data set (hereafter referred to as the unknown set) as possible [19,31]. Other attack scenarios are possible, but a marketer attack establishes an accepted baseline for risk. To simulate this attack scenario, we built a text-independent speaker identification model with a combination of x-vector extraction using Emphasized Channel Attention, Propagation, and Aggregation in Time-Delay Neural Network (ECAPA-TDNN) [32] and a downstream probabilistic linear discriminant analysis (PLDA)–based classifier [33,34], as described in detail in the following sections. Figure 1 shows the architecture of our model.

**Figure 1.** Speaker identification system architecture. During training, recordings from known speakers are fed into a pretrained speaker identification model (ECAPA-TDNN) to extract embeddings. These constitute a low-dimensional, latent representation for each recording that is enriched for speaker-identifying features (x-vectors). We used these x-vectors for known speakers to train a probabilistic linear discriminant analysis (PLDA) classifier and generate an average threshold for acceptance or rejection of a speaker match over several subsets. During testing, the extracted x-vectors are fed into the trained PLDA, and the training threshold is applied, resulting in a set of matches (or no matches) for each recording. ECAPA-TDNN: Emphasized Channel Attention, Propagation, and Aggregation in Time-Delay Neural Network.
Data

Overview

An ideal data set for our attack scenario would consist of (1) a set of elicited speech recordings from tasks typically used in clinical or research speech evaluations and (2) a set of unstructured speech recordings including the same speakers as in item 1 but acquired at a different time and place. This would allow us to directly assess the risk of reidentification of medical recordings by training a model on unstructured connected speech, such as what an adversary may find on the web. Such a data set does not exist. As such, we made use of 2 separate data sets. The first was a combination of the well-known VoxCeleb 1 and 2 data sets, which contain recordings from a web source of >7000 speakers [23,35,36]. The second was a medical speech data set from the Mayo Clinic, which contains recordings of commonly used elicited speech tasks but with fewer speakers.

VoxCeleb

The VoxCeleb 1 and 2 data sets are recent large-scale speaker identification data sets containing speech clips extracted from celebrity interviews on YouTube [23,35,36]. The utterances are examples of natural, real-world speech recorded under variable conditions from speakers of different ages, accents, and ethnicities. VoxCeleb 1 and 2 have a combined total of 1,281,762 recordings from 7363 speakers.

Mayo Clinic Speech Recordings

The Mayo Clinic clinical speech data set consists of recordings from elicited speech tasks in previously recorded speech assessments. Each speaker has a combination of clips from various tasks commonly used in a clinical speech evaluation, including sentence repetition, word repetition, paragraph reading, alternating motion rates (AMRs), sequential motion rates (SMRs), and vowel prolongation [11]. The clips from speakers vary in recording medium (cassette recording vs DVD), microphone distance, degree of background noise, and presence and severity of motor speech disorder or disorders. There are 19,195 recordings from 941 speakers (the breakdown is presented in Table 1).

<table>
<thead>
<tr>
<th>Table 1. Breakdown of number of recordings and speakers for each task in the Mayo Clinic clinical speech data set.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Vowel prolongation</td>
</tr>
<tr>
<td>&quot;Aaaaaah&quot;</td>
</tr>
<tr>
<td>AMR(^a)</td>
</tr>
<tr>
<td>SMR(^b)</td>
</tr>
<tr>
<td>Word repetition</td>
</tr>
<tr>
<td>&quot;Catapult&quot; and &quot;catastrophe&quot;</td>
</tr>
<tr>
<td>Sentence repetition</td>
</tr>
<tr>
<td>&quot;My physician...&quot;</td>
</tr>
<tr>
<td>Reading passage</td>
</tr>
</tbody>
</table>

\(^a\) AMR: alternating motion rate.

\(^b\) SMR: sequential motion rate.

\(^c\) 354 total unique speakers.

\(^d\) Samples instead of recordings.

\(^e\) 551 total unique speakers.

X-Vector Extraction Using ECAPA-TDNN

We generated speaker embeddings using a deep neural network to extract fixed-length embedding vectors (x-vectors) from speech recordings [32,34]. This technique has been shown to outperform previous embedding techniques such as i-vectors [37,38] while offering a competitive performance compared to newer end-to-end deep learning approaches [21,22]. Our network of choice was the state-of-the-art ECAPA-TDNN model, which was pretrained on a speaker identification task using VoxCeleb 1 and 2 [32]. This model extracts a 192-dimensional x-vector for each speech recording. The model is publicly available through SpeechBrain, an open-source artificial intelligence speech toolkit [39], and is hosted on Hugging Face.
PLDA Back-End Classifier

PLDA classifiers are a standard approach for speaker identification due to their ability to reliably extract speaker-specific information from an embedding space using both within- and between-speaker variance [33,40]. PLDA is a dimensionality reduction technique that projects data to a lower-dimensional space where different classes are maximally separated (ie, maximal between-class covariance). The advantage of PLDA over the standard linear discriminant analysis is that it can be generalized to unseen cases [41]. PLDA can then be used to determine whether 2 data points belong to the same class by projecting 2 data points to the latent space and using the distance between them as a measure of similarity. This works well for speaker identification as speaker embeddings are typically fed into a classifier in pairs, where the classifier’s role is to optimally reject or accept the hypothesis that the 2 recordings are from the same speaker. PLDA typically uses the log-likelihood ratio (probability of recordings belonging to the same class vs different classes) to measure similarity, commonly referred to as PLDA scores. During training of a PLDA classifier, PLDA scores for each pairwise comparison in the training set are computed and then used to set a threshold for determining potential speaker matches [33,40].

Our classifier was built and trained on a set of x-vectors extracted from either VoxCeleb or Mayo Clinic speech recordings using ECAPA-TDNN functions from SpeechBrain [39]. We aimed to maximize performance by giving the model multiple speech embeddings per speaker during training, each extracted from recordings under different degradation conditions (eg, varying background noise and microphone distances), which were then averaged to create a single speaker embedding [33].

Threshold Calculation for Acceptance or Rejection

During training, an optimal threshold needs to be determined to classify whether a given PLDA score represents a match, which can then be applied to new, unseen recordings. Matches that pass the threshold are then considered accepted matches. Generally, the equal error rate (EER) is used to select the threshold [21,22,24,33,34]. The use of the EER assumes that the cost of a false acceptance (FA) is the same as a false rejection (FR) such that the optimal threshold is 1, where the FA rate (FAR) equals the FR rate [22]. While this may be feasible for smaller data sets, when there are several million comparisons, the EER often generates many potential matches per speaker. As such, this can overwhelm the model early on and make it difficult for an adversary to find reliable matches. To scale up to large numbers of comparisons, the adversary must make decisions on how to calibrate the threshold calculation, such as penalizing FAs more heavily even if some true acceptances (TAs) are missed. From an adversary’s perspective, it is less costly to miss TAs if the identified accepted cases have a high likelihood of being true. In effect, precision is more important than recall. The detection cost function (equation 1 [42]) captures this well:

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

We take the cost of an FR ($C_{FR}$) multiplied by the total number of FRs and the prior probability of the target and add it to the cost of an FA ($C_{FA}$) multiplied by the total number of FAs and the complement of the prior probability.

Using this function, a threshold can be found by setting optimal cost and previous terms based on the adversary’s perspective (ie, avoiding FAs more aggressively) and then finding the FA and FR values that minimize the detection cost function (minDCF) [42]. For example, as the prior probability of the target is lowered (ie, if an adversary expects a small overlap), the calculation puts more emphasis on avoiding FAs (lower FAR) as compared to the EER. Increasing the cost of FAs and decreasing the cost of FRs further prevents FAs.

We used the minDCF with two parameter configurations: (1) the default configuration for the SpeechBrain implementation of the minDCF, where FAs and FRs are penalized equally ($C_{FR}$=1; $C_{FA}$=1; prior=0.01) [39]; and (2) a strict configuration with a higher penalty for FAs ($C_{FA}$=10; $C_{FR}$=0.1; prior=0.001).

Due to the large amount of training data in VoxCeleb, it was not computationally feasible to select a threshold for the entire set of identified speakers at once. In addition, we wanted to estimate thresholds that were representative of the population rather than any one subset of speakers. We used a bootstrap sampling technique in which we calculated a minDCF threshold on subsets of training speakers and averaged across runs to estimate the optimal threshold. For each run, latent representations from 2 random subsets of 100 speakers were selected from the training data and fed to the minDCF to calculate a threshold. If the 2 subsets had no overlapping speakers, the entire run was discarded as a threshold could not be calculated. We ran this process between 100 and 500 times depending on the overall number of speakers used for training the PLDA. Training with fewer speakers required fewer runs to converge on an optimal threshold.

Generating Experimental Speaker Sets

To model the attack scenario, we randomly sampled our data sets to generate the following speaker subsets:

1. **Known set:** this set represents speakers with identified audio data from a web source that the adversary has access to.
2. **Unknown-only set:** this set represents speakers in a shared data set who do not have identifiable audio on the web. No unknown-only speakers are present in the known set.
3. **Overlap set:** this set is a proxy for speakers in a shared data set who do have identifiable audio somewhere on the web. Some speakers from the known set are randomly selected to create this set.
4. **Unknown set:** this represents the full shared data set, consisting of both the unknown-only set and the overlap set.

The number of speakers per set varied based on the experiment. Furthermore, the number of speech recordings per speaker varied between the known and unknown sets. We used all available speech recordings per speaker in the known set but randomly selected only 1 recording per speaker in the unknown set. For overlapping speakers, the selected recording for the unknown set is a proxy for speakers in a shared data set.

---

We refer the reader to Wiepert et al. for a detailed explanation of the experiments and results. The minDCF is a widely used metric for evaluating speaker verification systems. It provides a principled way to select thresholds that balance the costs of misses and false alarms, taking into account the prior probabilities of the target speakers. The function captures the trade-off between precision and recall, which are critical considerations in speaker identification tasks. The minDCF formulation allows for flexible parameterization, enabling the optimization of cost functions that reflect the adversary’s perspective.
set was withheld from the known set. The limit of 1 sample per speaker in the unknown set was based on the nature of a supposed real-world data set where all speech is unlinked and partially deidentified, meaning that the adversary needs to separately find potential matches for each recording even if they come from the same speaker.

Because we randomly subsampled speakers to generate these sets, there is variation in the speakers selected for each experiment, which will result in variability in model performance that is dependent only on the data set. To account for this, we generated multiple speaker splits per experiment. The exact number of splits was dependent on the experiment.

Experiments

**VoxCeleb Realistic Experiments: Effect of Search Space Size**

We relied on VoxCeleb 1 and 2 to investigate the capability of an attack as a function of the size of the search space (i.e., the number of comparisons made to find matching speakers). We reidentified speakers by comparing each speaker in the known set to each speaker in the unknown set. Thus, the search space is the product of the sizes of the known and unknown sets. As such, an increase in either set will increase the number of comparisons. We considered both cases separately, which allowed us to consider one scenario that is dependent on the resources of the adversary (known set size) and another that is under the control of the sharing organization (unknown set size).

To construct a realistic scenario, we assumed that the known and unknown sets would have a low degree of speaker overlap. To justify this assumption, one can consider what would be involved in constructing a set of known speakers. In the absence of metadata about the unknown speakers (e.g., the ages and location), there would be no way for an adversary to target a specific population to build their known set. It is unlikely to be feasible for an adversary to manually collect and label speech recordings for a large proportion of the population. Instead, an adversary would likely need to rely on a programmatic approach using easily accessible identifiable audio, such as scraping audio from social media and video- or audio-sharing websites [43]. It is worth noting that this would still be difficult because of several confounding factors: (1) not all members of the population use these websites; (2) not all users have publicly accessible accounts; (3) users with publicly accessible accounts may not have identifiable information linked to them; (4) some accounts post audio or video from multiple speakers, including speakers who also have their own accounts; (5) many users do not post at all; and (6) the population of users is not representative of the general US population, let alone the subset with speech disorders—in terms of the distribution of both age and geographic area [44]. As such, there is no reason to suspect that a patient in a shared medical speech data set would have a high likelihood of existing in an adversary’s set of identified audio recordings.

We also assumed that the adversary would not know which unknown speakers, if any, exist in the known speaker set. Therefore, the adversary must consider all potential matches rather than only focusing on the \( N \) overall best matches, where \( N \) is the known overlap. This would reduce the reliability of any match because the likelihood of all potential matches being true is lower than the likelihood of the best \( N \) matches being true.

We first trained the speaker identification model with the number of speakers in the known set increasing from 1000 to 7205 while maintaining a static unknown set size of 163 speakers, with low speaker overlap between sets (\( n=5, 3.1\% \) speakers in the overlap set and \( n=158, 96.9\% \) in the unknown-only set).

We then trained the model with a fixed known set size of 6000 speakers while increasing the number of speakers in the unknown set from 150 to 1000 speakers and maintaining a low overlap of 5 speakers.

Given the low number of overlapping speakers and overall large set sizes, we generated 50 speaker splits for each set size of interest (known set: 1000, 4000, and 7205; unknown set: 150, 500, and 1000).

The acceptance threshold for these experiments was set using the strict minDCF configuration. Experimental parameters are summarized in Table 2.
Table 2. Experimental parameters, including number of runs; set sizes; and minimum detection cost function (minDCF) parameters such as the cost of a false acceptance (CFA), cost of a false rejection (CFR), and prior probability (prior).

<table>
<thead>
<tr>
<th>Experiment</th>
<th>Runs, n</th>
<th>Set size, total speakers</th>
<th>minDCF parameters</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Known</td>
<td>Unknown</td>
</tr>
<tr>
<td>VoxCeleb: effect of search space size and known-overlap worst-case scenario</td>
<td>50</td>
<td>1000 to 7205 (varied known)</td>
<td>163 (varied known)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>6000 (varied unknown)</td>
<td>150 to 1000 (varied unknown)</td>
</tr>
<tr>
<td>VoxCeleb: full-overlap worst-case scenario</td>
<td>20</td>
<td>1000 to 7205 (varied known)</td>
<td>163 (varied known)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>6000 (varied unknown)</td>
<td>150 to 1000 (varied unknown)</td>
</tr>
<tr>
<td>Mayo Clinic speech recordings: cross-task</td>
<td>20</td>
<td>500</td>
<td>55</td>
</tr>
<tr>
<td>Mayo Clinic speech recordings: within task</td>
<td>20</td>
<td>500^a</td>
<td>55</td>
</tr>
</tbody>
</table>

^Word repetition: 299 speakers; reading passage: 466 speakers.

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

There are two important initial assumptions in our construction of realistic experiments: (1) the adversary was unaware of the amount of overlap between known and unknown sets, and (2) the amount of overlap was low. Thus, we considered how reidentification risk would be affected if either assumption was incorrect.

First, we considered a potential worst-case scenario in which the adversary did know the number of overlap speakers $N$ and, therefore, was able to limit potential matches to the top $N$ best matches. As previously mentioned, limiting the number of matches could theoretically improve model reliability, and further reducing the number of matches could produce more noticeable effects. We leveraged our base results from the realistic experiments and only considered the top $N$ best matches.

Next, we considered a less realistic worst-case scenario in which all unknown speakers exist in the known speaker set. From an adversary’s perspective, a full-overlap scenario would provide the best chance for them to successfully reidentify speakers because most FAs occur when the model finds a match for unknown speakers who are not in the known speaker set.

We assessed this scenario by replicating the realistic experiments with full overlap between the known and unknown sets. That is, regardless of the unknown set size, all speakers also exist in the known set (no unknown-only set). When increasing the known set size with a fixed unknown set of 163 speakers, the overlap set consists of all 163 speakers, and when increasing the unknown set size with a fixed unknown set, the overlap set is the same as the unknown set size of interest (150, 500, and 1000). In this scenario, we generated only 20 speaker splits for each set size of interest as the larger overlap set led to less variance across runs.

As in the realistic experiments, the acceptance threshold was set using the strict minDCF configuration. Experimental parameters are summarized in Table 2.

**Mayo Clinic Speech Recording Experiments: Effect of Speech Task**

Next, we shifted our focus from the public VoxCeleb data set to a private data set of Mayo Clinic medical speech recordings to look at factors specific to a clinical speech data set, such as whether certain elicited tasks are easier for reidentification and whether certain tasks are easier to link recordings to the same speaker across tasks (pooling) increases risk.

We first compared the performance of the speaker identification model across the various elicited speech tasks in the Mayo Clinic data set based on the same adversarial attack scenario used with the VoxCeleb experiments. In this scenario, the cross-task performance aligns with a real-world case in which the training data contain connected speech recordings (ie, recordings of continuous sequences of sounds such as those of spoken language) but speakers are reidentified using a variety of elicited speech tasks (Table 1). Each task has a different degree of similarity to connected speech (left: most; right: least):

Reading passage > sentence repetition > word repetition > SMR > AMR > vowel prolongation

The reading passage is essentially real-world connected speech in terms of content and duration, but sentence repetition is closer to the connected speech seen in most speech data sets [23]. As such, we selected sentence repetition recordings for speakers in the known set.

The resulting known set comprised 500 speakers and included all sentence repetition recordings, excluding any repetitions of
the physician sentence (“My physician wrote out a prescription”), which was saved for the unknown set. We then generated separate unknown sets for each elicited task with 55 speakers (n=5, 9% overlap and n=50, 91% unknown only) who had both sentence repetition recordings and a recording for the given reidentification task (eg, “My physician...” sentence and AMRs).

The known and unknown set sizes were bounded by the number of speakers with sentence repetition recordings (587 speakers) as the sentence-sentence configuration required enough speakers to create a separate known and unknown-only set. We also considered the sentence-sentence configuration (ie, sentence repetitions in both the known and unknown sets) as the realistic baseline.

As a secondary part of this experiment, we pooled all available recordings from all elicited speech tasks (by averaging their embeddings) to generate an unknown set in which the adversary could link recordings from a given speaker (ie, there would be more speech for each unknown speaker).

In addition to the cross-task performance, we compared the within-task performance—where the same elicited speech task is used for both known and unknown speakers—to determine whether anything about the nature of a given speech task affected reidentification. For example, the variance across recordings for the sentence repetition task reflects a combination of static speaker factors (eg, identity and age), dynamic speaker factors (prosody, eg, the same speaker may emphasize different words in a sentence on repeated trials), and content factors (ie, different words in different sentences). In contrast, a task such as AMR involves repeating the same syllable as regularly and rapidly as possible, with most of the variance across speakers likely resulting from static speaker factors. A priori, considering all the elicited tasks, one would expect the proportion of variance across speakers due to dynamic speaker factors to decrease following the same scale as similarity to natural speech. The reading passage would have the most variance due to dynamic speaker factors alone, whereas vowel prolongation would have the least variance. By removing the confounding variable of different elicited tasks for known and unknown speakers (ie, the model is both trained and tested on the same task), we can ascertain whether the qualities of the speech task itself influence reidentification.

We used the same set sizes as the cross-task experiments (500 known, 55 unknown, and 5 overlap) but used recordings from the same elicited speech task in both the known and unknown sets. This setup required at least 2 recordings per speaker for each task. Some tasks had <500 unique speakers or not enough recordings (word repetition and reading passage), so not every known set had exactly 500 speakers. The word repetition task had 299 speakers, and the reading passage task had 466 speakers.

To account for the decrease in the amount of data as compared to the VoxCeleb experiments, we generated only 20 speaker splits per task with default minDCF parameters. Experimental parameters are summarized in Table 2.

### Statistical Analyses

Given that we were simulating an adversarial attack and not optimizing a model, we used random splitting to account for the potential of outlier cases, wherein specific configurations of speakers in the known and unknown sets had a higher-than-average risk of reidentification. We first randomly sampled our larger data set either 20 or 50 times depending on the experiment to generate speaker splits (known, unknown, and overlap sets). We also randomly selected a single recording per speaker in the unknown set to mitigate utterance effects. Furthermore, we used bootstrap sampling of the known (training) set to estimate our acceptance threshold by feeding cohorts of 100 speakers to the minDCF function between 100 and 500 times to converge on an optimal threshold. The exact number of runs was dependent on the overall number of speakers in the known set.

Our primary outcome of interest was the average number of FAs, where the model accepts a match for an unknown speaker without a true match, compared to TAs over several subsampled data sets. Using these counts, we also calculated precision. These metrics informed the reliability of reidentification. Note that TAs and FAs are functionally equivalent to true and false positives, respectively. Using the counts, we also calculated the Pearson correlation coefficient between FAs and set size along with the FAR to determine whether a linear correlation existed between the number of FAs and the number of speakers or comparisons. A 2 tailed t test was performed to determine the significance of each correlation.

### Ethical Considerations

The primary data type for this work was clipped speech recordings from either VoxCeleb or our Mayo Clinic clinical speech data set. We could not deidentify the data due to the nature of our work, and the data sets were not anonymous. The VoxCeleb data set has no privacy protections or additional consent processes in place given its public nature—all recordings come from interviews of celebrities posted on YouTube [23,35,36]. For the Mayo Clinic clinical speech data set, we submitted an institutional review board application to the Mayo Clinic to gain permission to use the data. Our work was deemed exempt from additional consent requirements and granted a waiver of HIPAA authorization considering the secondary nature of the analysis. No compensation was offered to participants in the original studies. As the clinical data set may contain private health information, we do not share any recordings or models trained on the clinical recordings. Only researchers at our institution with proper permission can access the clinical data set.

### Results

#### VoxCeleb Realistic Experiments: Effect of Search Space Size

When training the speaker identification model with increasing numbers of speakers in the known set while maintaining a static unknown set size with low speaker overlap between sets, we found that increasing the number of speakers in the known set resulted in an increase in the mean number of FAs while TAs
remained stable, with a linear correlation between FAs and the number of known speakers ($r=0.30$; $P<.001$; $t_{148}=3.89$; Figure 2A). Increasing the size of the unknown set had a similar yet more pronounced effect than increasing the known set size, with a higher linear correlation between FAs and the number of unknown speakers ($r=0.60$; $P<.001$; $t_{148}=9.21$; Figure 2B).

The difference in effect can be understood based on the geometry of the search space. While the unknown set remains substantially smaller than the known set, adding a speaker to the unknown set will result in a larger increase in the search space than adding a speaker to the known set. As such, we can better demonstrate the overall trend in FAs by considering the results in terms of total comparisons (i.e., search space size) rather than individual set size.

**Figure 2.** Number of true acceptances (TAs) and false acceptances (FAs) for the speaker recognition model in a realistic scenario using VoxCeleb. (A) shows the counts when varying the number of known speakers while keeping the number of unknown speakers static, (B) shows the counts when varying the number of unknown speakers while keeping the number of known speakers static, and (C) shows the overall trend in terms of the number of comparisons made (i.e., the search space size=known×unknown speakers). All plots (A-C) include the Pearson correlation coefficient and corresponding significance for FAs and number of speakers or comparisons. Each run is plotted as a single circle, with red horizontal lines indicating the mean number of FAs and green horizontal lines indicating the mean number of TAs. minDCF: minimum detection cost function.

We observed that there was a high positive linear correlation between FAs and the number of comparisons ($r=0.69$; $P<.001$; $t_{198}=13.54$; Figure 2C), with the mean FAs increasing from 0.04 to 2.84 while TAs remained stable. The ratio between FA and TA (FA/TA) rose from 0.02 at $1 \times 10^5$ comparisons to 1.41 at $6 \times 10^6$ comparisons, with a near 1:1 ratio at the midpoint of $3 \times 10^6$ comparisons. There was a corresponding drop in precision (Figure 3A). It was notable that the FAR remained low and...
relatively stable, averaging at $4.152 \times 10^{-7}$ (SD $7.255 \times 10^{-7}$; Figure 3B), indicating that the demonstrated trend should hold for the larger numbers of comparisons that we would expect to see in a real attack.

We further observed that using a stricter threshold for matches resulted in our model selecting only 1 match per speaker. This is functionally the same as limiting matches to only the best potential match for each speaker (rank-1 matches), which is an option for an adversary to increase reliability without knowledge of the amount of overlap.

Figure 3. Precision and false acceptance rates (FARs) for the speaker recognition model in a realistic scenario using VoxCeleb. Precision (A) and FARs (B) are shown as a function of the number of comparisons. For both plots, each run is represented by a circle, and the mean is represented by a horizontal black line. FA: false acceptance; minDCF: minimum detection cost function; TA: true acceptance.

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

When only considering the top $N$ best matches, we found that there was still a trend of increasing FAs, with a high linear correlation with the number of comparisons ($r=0.70; P<.001; t_{198}=13.72$; Figure 4A). The FA/TA ratio increased from 0.02 at $1 \times 10^5$ comparisons to 1.24 at $6 \times 10^6$ comparisons and again had a near 1:1 ratio at $3 \times 10^6$ comparisons. These results indicate that some FAs were seen as better matches than some TAs, as further supported by the associated drop in precision (Figure 4B).

When all unknown speakers existed in the known speaker set, the performance improved significantly, with most matches being correct (Figure 4C). Even so, there was still a high positive linear trend for FAs, indicating that, at high overlap, some FAs were ranked higher than TAs ($r=0.67; P<.001; t_{78}=7.98$; Figure 4D). The FA/TA ratio exhibited a fairly large increase considering the number of TAs, increasing from 0.0008 at $1 \times 10^5$ comparisons to 0.008 at $6 \times 10^6$ comparisons. This is surprising given that, for the realistic experiments, all FAs were associated with matches for nonoverlapping speakers.
Figure 4. Results for our speaker recognition model in worst-case scenarios using VoxCeleb. (A) shows the true acceptance (TA) and false acceptance (FA) counts for a known-overlap scenario (limited to N=5 best matches), whereas (B) shows the corresponding precision as a function of the number of comparisons (search space size). (C) and (D) show the FA and TA counts for a full-overlap scenario in which all unknown speakers are present in the known speaker set as a function of the number of comparisons (search space size). (A) and (C) also show the Pearson correlation coefficient and corresponding significance between FAs and number of comparisons. Each run is plotted as a single circle, with red horizontal lines indicating the mean number of FAs, green horizontal lines indicating the mean number of TAs, and black horizontal lines indicating the mean precision. minDCF: minimum detection cost function.

Mayo Clinic Speech Recording Experiments: Effect of Speech Task

We first compared the performance of the speaker identification model across the various elicited speech tasks in the Mayo Clinic data set based on the same adversarial attack scenario used in the VoxCeleb experiments. We observed that the total number of acceptances decreased as the unknown speaker tasks became less similar to the known speaker task, but the proportion of TAs and FAs also varied. This made it more difficult to determine the performance through counts alone (Figure 5A). When considering precision and FA/TA ratio instead, we found that the baseline (sentence-sentence) had the best performance, although the average precision was not high (FA/TA=0.54; precision=66.5%; Figure 5B). The paragraph reading, word repetition, and SMR tasks had a worse performance than the baseline but were comparable to each other in terms of both precision (Figure 5B) and FA/TA ratios (reading passage: FA/TA=1.09; word repetition: FA/TA=0.72; SMR: FA/TA=0.85). However, the AMR and vowel prolongation tasks had extremely low precision and high FA/TA ratios. Vowel prolongation, in particular, had a precision of 0 (almost no TAs across runs) but a high number of FAs, resulting in a ratio of 98.5. Pooling resulted in decreased performance compared to
the baseline and the top-performing tasks in terms of both precision (approximately 36%) and FA/TA ratio (2.56). This was likely due to the influence of AMR and vowel prolongation recordings.

The within-task results did not exhibit the same effect as the cross-task results. We found that all tasks reidentified the overlapping speakers (TA=10) but the number of FAs varied drastically across tasks (Figure 5C). Previously, the baseline had the best performance, whereas we instead observed that the SMR and vowel prolongation tasks had the highest precision (Figure 5D), as well as FA/TA ratios of 0.35 and 0.39, respectively. In fact, as tasks became more dissimilar from connected speech and had less variance due to dynamic speaker factors, they saw a relative increase in performance compared to the cross-task scenario. Word repetition was the only exception to this, with lower precision and a greater FA/TA ratio of 2.02 as compared to the cross-task performance.

Figure 5. Results for our speaker recognition model using the Mayo Clinic clinical speech data set. (A) and (B) show cross-task results, in which recordings for known speakers are always sentence repetition but the task for unknown speaker recordings varies. The baseline is when sentence repetitions are in both the known and unknown sets. Pooling is when all recordings for an unknown speaker are linked together across all tasks. (A) shows the breakdown of counts for this case, whereas (B) is the corresponding precision. (C) and (D) show within-task results, where tasks for known and unknown speakers are always the same. (C) is the breakdown of counts for this case, whereas (D) is the corresponding precision. Each run is plotted as a single circle, with red horizontal lines indicating the mean number of false acceptances (FAs), green horizontal lines indicating the mean number of true acceptances (TAs), and black horizontal lines indicating the mean precision. AMR: alternating motion rate; minDCF: minimum detection cost function; SMR: sequential motion rate.
Discussion

Principal Findings

In this study, we investigated the risk of reidentification of unidentified speech recordings without any other speaker- or recording-related metadata. To do so, we performed a series of experiments reflecting a marketer attack by an adversary with access to identified recordings from a large set of speakers and the capability to train a speaker identification model, which would then be used to reidentify unknown speakers in a shared data set. We systematically considered how changes in the size of the data sets and the nature of the speech recordings affected the risk of reidentification. We found that it is feasible to use a speaker identification design—a deep learning speaker embedding extractor (x-vectors) coupled with a PLDA back end—to reidentify speakers in an unknown set of recordings by matching them to recordings from a set of known speakers. Given the performance of current state-of-the-art speaker identification models, this is not surprising. However, these models have only rarely been applied in an adversarial attack scenario [24,25] (ie, their potential as an attack tool for an adversary who aims to reidentify speakers in a shared or publicly available data set was largely unknown). Furthermore, the feasibility of such an attack has not been considered and may have been assumed to be low for speech recordings stripped of all metadata (sometimes referred to as deidentified or anonymous in the literature) without considering the identifiability of the acoustic signal itself [45-48].

Our findings suggest that this is not true. Consistent with a previous study that found a high reidentification risk for an unknown speaker with known sets of up to 250 speakers (search space of ≤250 comparisons) [25], we observed that risk was indeed high for small search spaces. For example, when attempting to reidentify 5 overlapping speakers between a small set of unknown speakers (n=163) and a moderate set of known speakers (n=1000), our model had nearly perfect precision (Figure 3A) and identified 2 speakers on average (FA/TA=0.02; Figure 2A). However, our experiments allowed us to extend this to more realistic search spaces, such as scenarios in which an adversary uses a known speaker set of up to 7205 speakers and an unknown speaker set of up to 1000 speakers (search space of ≤56 million comparisons). We observed that the risk dropped sharply as the search space grew. The FAR was relatively stable at 4.152 \times 10^{-7} (Figure 3B), which translates to an average increase of 1 FA for every 2.5 million comparisons. This is a key take-home message from these experiments—increasing the size of the search space, whether by increasing the size of the adversary’s set of identified recordings or of the shared data set, resulted in a corresponding increase in the number of FAs. Given that the number of overlapping speakers remained constant, this suggests that the primary driver of FAs is the size of the nonoverlapping known-to-unknown comparison space (ie, most FAs arise from nonoverlapping unknown speakers being falsely matched to known speakers). In fact, all FAs in the realistic experiments corresponded to nonoverlapping unknown speakers. Here, it is worth noting that, in the experiments in which we only considered the top N matches (where N=number of overlapping speakers), this trend remained true because some of the FAs scored higher than TAs (Figure 4). This suggests that for a sufficiently large search space, even considering only the best N matches will result in many FAs. We pushed this line of reasoning to its limit by considering a worst-case scenario of full overlap in which all unknown speakers had a true match. Even in this scenario, there were still many FAs, and the proportion of FAs increased with increasing search space size. Importantly, this scenario showed that overlapping speakers can still be falsely matched when the overlap is high.

Our experiments with the Mayo Clinic clinical speech recordings allowed us to assess the influence of speech task based on both cross-task and within-task performance. When the model was trained on sentence repetition (ie, the known data set consisted of sentence recordings) and then applied to other tasks (ie, the unknown set consisted of elicited, nonsentence speech), all tasks performed below the baseline, but performance deteriorated most drastically for the less connected speech–like tasks such as AMR and vowel prolongation. These results can be understood with reference to the default minDCF settings, which would penalize FAs and FRs equally. The threshold was chosen using sentence repetition task recordings such that, in most instances, all overlapping speakers were reidentified for unknown sets with connected speech tasks (sentence repetition, paragraph reading, word repetition, and SMRs). The minDCF threshold for these similar tasks resulted in fewer overall acceptances (higher FR rate), but as the tasks diverged from sentence repetition with respect to the degree of connectedness, they were also less likely to be FAs. This suggests that identifiable characteristics learned from training on the sentence repetition task translate well to other connected speech tasks. It also demonstrates the difficulty of choosing a threshold when the tasks in the known set are different from those in the unknown set. Because of the differences within a speaker across tasks, it becomes hard to balance TAs with the flood of FAs as the search space increases. In this instance, a slightly stricter threshold may have been better for the adversary. In contrast, the non–connected speech tasks (AMRs and vowel prolongation) had almost no TAs and a high number of FAs, suggesting that identifiable characteristics from connected speech tasks do not translate to non–connected speech tasks. This is not unexpected given that models perform worse when tested on data that are dissimilar from the training data [49,50]. Following this, we also found that pooling across tasks decreased performance from the baseline. Generally, having more data for a speaker is expected to improve performance, but it is possible that adding recordings of nonsentence tasks to the unknown set hurt performance because the identifiable characteristics are different across tasks and the system is unable to accommodate them. In other words, any helpful characteristics from the connected speech tasks were cancelled out by competing characteristics from the non–connected speech tasks.

In the within-task scenarios, where the known and unknown sets were made up of the same task, the reidentification power for overlapping speakers was better than in the cross-task scenario, but the tasks exhibited vastly different FA rates. In fact, many tasks that were different from connected speech saw improved performance. For example, vowel prolongation, which
is nonconnected and the most perceptually different from sentence repetition, exhibited the worst cross-task performance but the second-best within-task performance. This may be because less connected tasks have fewer interfering dynamic speaker factors such that they isolate well the acoustic features that are tied to identity.

Another important finding is that performance for sentence repetition was much weaker than expected based on the VoxCeleb experiments with a larger number of comparisons. We suspect that this may be due to a combination of factors. First, it may be more difficult to differentiate speakers in an unknown set of elicited recordings in which every speaker utters the same sentence. Second, the clinical recordings were all made by patients referred for a speech examination. Consequently, the resulting cohort contained mostly speech with abnormalities, which may impact the PLDA performance. Third, the Mayo Clinic clinical speech data set is smaller than the VoxCeleb data set in terms of both the number of speakers and the number of recordings per speaker, and the recordings are also shorter in duration. This likely had a negative impact on the training of the PLDA classification back end. It remains unknown whether larger clinical data sets or data sets with more recordings per speaker may yield findings more similar to the VoxCeleb results.

Taken together, our findings suggest that the risk of reidentification for a set of clinical speech recordings devoid of any metadata in an attack scenario such as the one we considered in this study is influenced by (1) the number of comparisons that an adversary must consider, which is a function of the size of both the unknown and known data sets; (2) the similarity between the tasks or recordings in the unknown and known data sets; and (3) the characteristics of the recordings in the unknown data set, such as degree of speaker variance and presence and type of speech disorders. These findings translate to actionable goals for both an adversary and the sharing organization.

Mitigating Privacy Risk

While we assumed that the sharing organization had already reduced risk by stripping recordings of demographic (eg, age or gender) or recording (eg, date or location) metadata, we additionally suggest that reidentification risks could be further reduced by increasing the search space (ie, larger shared data set size) or decreasing the similarity between shared recordings and publicly available recordings (eg, sharing vowel prolongation recordings as long as a publicly available vowel prolongation recording data set does not exist or sharing a larger variety of speech disorder recordings instead of those for a single disorder). Even if the number of overlapping speakers increased with the size of the shared data set, the results from the full-overlap scenario indicate that a model could still have reduced reliability due to an increasing FAR.

In contrast, an adversary can also use this knowledge to enhance their attacks. From their perspective, any additional information that can reduce the search space or increase the similarity between recordings will increase the reliability of speaker matches. This could involve using demographics such as gender, be they shared or predicted by a separate model, to rapidly reduce the number of comparisons. For instance, when the gender balance is 50:50, comparing unknown male individuals to known male individuals would reduce the number of comparisons by 75% (eg, from 6 million to 1.5 million). The adversary may also seek out publicly available recordings of speech with abnormalities to refine their model or models or reduce the search space based on speech disorders. If social media groups exist where identified users with certain medical or speech disorders post videos or audio, an adversary could restrict their known set to these users. Similarly, research participants and support staff may also influence risk through disclosure of participation. By disclosing participation in a study known to share speech recordings, a participant would effectively reduce the size of the known set to 1, increasing their individual risk of reidentification. In addition, having a confirmed match can increase risk overall as the adversary would have a baseline to determine the reliability of matches [51]. Although the focus of this investigation was on the change in relative risk with changes in data set size and speech task, it is worth considering our findings in the context of other factors that impact risk in practice. The most obvious factor is the availability of additional metadata on the speakers or recording. In this respect, it is worth noting that sufficient demographic data, even in the absence of speech, are well known to carry a significant risk of reidentification [19,52]. If any aspect of the metadata makes a patient population unique (ie, there is only one person in a given age range), the risk of reidentification increases [12,14]. Furthermore, the risk is not necessarily the same for all speakers or groups. For example, individuals with rare speech disorders, accents, or other qualities may be easier to match across known and unknown data sets. There may also be identifiable content in the recordings. During less structured speech tasks such as recordings of open-ended conversations, participants may disclose identifiable information about themselves (eg, participants saying where they live). Removing these spoken identifiers is an active area of research [25].

However, it is important to acknowledge that simply because records are vulnerable to reidentification does not mean that they would be reidentified. Notably, when assessing privacy concerns, the probability of reidentification during an attack is conditional on the probability of an attack occurring in the first place [52]. In most instances in which data are shared, the receiving organization or individual will not have any incentive to attempt reidentification. The sharing organization and, in some cases, a receiving organization may also take steps to discourage the risk of an attack. These may take the form of legal (eg, data-sharing agreements) or technical (eg, limited, monitored access) deterrents to a reidentification attack [53]. In contrast, the risk of an attack may be higher for publicly available data sets [54], but there may also be a greater risk of reidentification without a targeted attack. For example, in the field of facial recognition, some companies have scraped billions of photos from publicly available websites to create massive databases with tens of millions of unique faces. These are then used to train a matching algorithm [43], which an end user could query using a photo of an unknown face and obtain a ranked list of matching faces and the source (eg, Facebook). The end user can visit the source website and instantly gain access to other data that may increase or decrease their confidence in a match as well as provide feedback on matches, thereby gradually

https://ai.jmir.org/2024/1/e52054
increasing the performance of the tool as well as the number of known faces. If similar databases are built for speech recordings, they will certainly include publicly available medical speech recordings. Every query to the model would then represent a threat to such a public sample being matched to a queried recording regardless of the intent of the user who queried the model. Such a scenario is difficult to simulate because of the continuously improving nature of the algorithm and the fact that users would incorporate various degrees of nonspeech data.

Refining from publicly releasing data sets is an obvious mitigation strategy for some of these threats. However, the risk of reidentification must always be balanced with the benefit of data sharing as larger, more representative data sets for the development and testing of digital tools may benefit patients. It is critical that policy makers consider this balance in the context of the rapidly evolving field of artificial intelligence. Naive approaches such as the “deidentification release-and-forget model” are unlikely to provide sufficient protection [55]. Similarly, informed consent for public release is problematic because the risk of reidentification will be neither static nor easily quantifiable over time. This has led to the development of potential alternative approaches, such as data trusts, synthetic data, federated learning, and secure multiparty computation [56-59].

Limitations

It should be recognized that there are several notable limitations to our investigation. First, while we relied on state-of-the-art learning architectures, the risk may differ if other computational approaches are considered [21,22]. Second, we did not consider multistage adversarial attacks in which one model is used to predict a demographic, such as sex or age, which is then used to limit the search space, or a scenario in which an adversary manually goes through all potential matches to attempt manual identity verification. However, such approaches would introduce additional uncertainty for the adversary as they would generate predictions for an out-of-sample data set of speech with abnormalities, meaning that accuracy may be lower than expected and the resulting filtered data set may still require many comparisons, in which case our results would apply [60,61]. Third, we did not directly consider the risk of healthy speech versus speech with abnormalities. Nearly all recordings in the Mayo Clinic speech data set contain speech with abnormalities, whereas all VoxCeleb recordings are from healthy speakers. Ideally, there would be a single data set containing both. Fourth, it should be noted that, beyond methodological limitations, our results may not generalize well outside of the United States as the VoxCeleb data have a strong US bias and all the Mayo Clinic recordings were captured in the United States. As such, it will be important to conduct future experiments that leverage alternative computational architectures, more complex adversarial attacks, conversational speech, and data from other geographic regions to assess the reidentification risk for medical speech data more comprehensively.

In addition, there is an important implication of the VoxCeleb experimental design. As we were interested in a range of set sizes and wanted to complete multiple runs for each size, we combined the train and validation sets from VoxCeleb 1 and 2 and randomly selected a holdout set. However, the ECAPA-TDNN model used for extracting embeddings was pretrained on VoxCeleb, meaning it was exposed to most of the recordings (ie, all but the validation cases) during the original training step [32]. The embeddings are almost certainly superior to what one may have obtained if the embedding model was retrained for each of our splits. Unfortunately, that is not a computationally feasible experimental design. Furthermore, superior embeddings mean we are likely to overestimate risk and draw more conservative conclusions. Given the stakes—reidentification of anonymous research patients—we feel this decision was justified. We also ran a set of experiments using the VoxCeleb validation set as our unknown set (Multimedia Appendix 1). This only allowed for a small unknown set with fixed speakers across runs, so it may be overly optimistic regarding risk. In our opinion, the true risk lies in between our main results and the supplementary results.

Conclusions

In summary, our findings suggest that while the acoustic signal alone can be used for reidentification, the practical risk of reidentification for speech recordings, including elicited recordings typically captured as part of a medical speech examination, is low with sufficiently large search spaces. This risk does vary based on the exact size of the search space—which is dependent on the number of speakers in the known and unknown sets—as well as the similarity of the speech tasks in each set. This provides actionable recommendations to further increase participant privacy and considerations for policy regarding the public release of speech recordings. Finally, we also provide ideas for future studies to extend this work, most notably the need to assess other model architectures and data sets as improvements in speaker identification could substantially increase reidentification risk.

Data Availability

The VoxCeleb 1 and 2 data sets analyzed during this study are available in the VoxCeleb repository [62]. Our Mayo Clinic clinical speech recordings data set analyzed during this study is not publicly available due to the privacy risks related to the release of clinical speech data and are not available by request. We used Python (Python Software Foundation) to implement our code for preprocessing, extracting speaker embeddings, generating subsampled data sets, and running the probabilistic linear discriminant analysis. The source code is available on the internet [63]. The repository also contains detailed documentation for using the scripts.

Acknowledgments

No generative language models were used when writing the manuscript.
Authors' Contributions

DW, BAM, and HB conceived the ideas presented in this study and validated the results. JRD, RLU, and DTJ provided the necessary resources for this study. DW, JLS, and HB developed the methodology for the experiments. DW curated the data for the Mayo Clinic speech recording data set. DW and HB developed the code for running the experiments and visualizing the results. DW conducted formal statistical analysis of the data. DW and HB wrote the original draft of the manuscript. All authors have reviewed and edited the manuscript. DTJ and HB supervised.

Conflicts of Interest

BAM, JRD, RLU, JLS, DTJ, and HB receive funding from the National Institutes of Health. All other authors declare no other conflicts of interest.

Multimedia Appendix 1
An additional set of experiments using only the VoxCeleb validation set as the unknown set. This only allowed for a small unknown set with fixed speakers across runs, so it may be overly optimistic regarding risk. These experiments define a lower bound for risk as compared to the original experiments that draw more conservative conclusions and may overestimate risk.

[DOCX File, 245 KB - ai_v3s1e52054_app1.docx ]

References


61. Wiepert et alJMIR AI 2024 | vol. 3 | e52054 | p.126https://ai.jmir.org/2024/1/e52054

XSL-FO RenderX

VoxCeleb: a large scale audio-visual dataset of human speech. Visual Geometry Group. URL: https://www.robots.ox.ac.uk/~vgg/data/voxceleb/ [accessed 2024-03-05]


Abbreviations
- AMR: alternating motion rate
- ECAPA-TDNN: Emphasized Channel Attention, Propagation, and Aggregation in Time-Delay Neural Network
- EER: equal error rate
- FA: false acceptance
- FAR: false acceptance rate
- FR: false rejection
- HIPAA: Health Insurance Portability and Accountability Act
- minDCF: minimum detection cost function
- PLDA: probabilistic linear discriminant analysis
- SMR: sequential motion rate
- TA: true acceptance

©Daniela Wiepert, Bradley A Malin, Joseph R Duffy, Rene L Utianski, John L Stricker, David T Jones, Hugo Botha. Originally published in JMIR AI (https://ai.jmir.org), 15.03.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR AI, is properly cited. The complete bibliographic information, a link to the original publication on https://www.ai.jmir.org/, as well as this copyright and license information must be included.
Privacy-Preserving Federated Survival Support Vector Machines for Cross-Institutional Time-To-Event Analysis: Algorithm Development and Validation

Julian Späth¹, MSc; Zeno Sewald², BSc; Niklas Probul¹, MSc; Magali Berland³, PhD; Mathieu Almeida³, PhD; Nicolas Pons³, PhD; Emmanuelle Le Chatelier³, PhD; Pere Ginès⁴,⁵,⁶, MD, PhD; Cristina Solé⁴,⁵,⁶, MD; Adrià Juanola⁴,⁵,², MD, PhD; Josch Pauling², PhD; Jan Baumbach¹, Prof Dr

¹Institute for Computational Systems Biology, University of Hamburg, Hamburg, Germany
²LipiTUM, Chair of Experimental Bioinformatics, TUM School of Life Sciences, Technical University of Munich, Freising, Germany
³MetaGenoPolis, INRAE, Université Paris-Saclay, Jouy-en-Josas, France
⁴Liver Unit, Hospital Clínic de Barcelona, Barcelona, Spain
⁵Institut d’Investigacions Biomèdiques August Pi i Sunyer (IDIBAPS), Barcelona, Spain
⁶Centro de Investigacion en Red de Enfermedades hepaticas y Digestivas (CIBERehD), Madrid, Spain
⁷Faculty of Medicine and Health Sciences, University of Barcelona, Barcelona, Spain

Corresponding Author:
Julian Späth, MSc
Institute for Computational Systems Biology
University of Hamburg
Notkestrasse 9
Hamburg, 22607
Germany
Phone: 49 15750665331
Email: julian.alexander.spaeth@uni-hamburg.de

Abstract

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

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

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

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

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

(JMIR AI 2024;3:e47652) doi:10.2196/47652
Introduction

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

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

Especially in combination with privacy-enhancing techniques (PETs), model parameters can be exchanged securely, such that a global aggregator or potential attacker cannot even see the local parameters of each participant [11]. This secure exchange of model parameters is necessary to comply with the GDPR, as even local models can be considered personal data [12]. Therefore, FL enables the training on a significantly larger data set compared with single-institution scenarios. While federated algorithms still often struggle with communication efficiency, the significantly increased amount of data can offset this performance issue, making FL a serious competitor to classical ML. Additionally, since FL models are trained on a larger variety of data, they typically generalize better than traditional ML models and even generalize faster in some cases [13,14]. Many FL approaches are already published for biomedical applications, such as medical imaging analysis, genome-wide association studies, or gene expression analysis [15-17].

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

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

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

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

Federated Survival SVM

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

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

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

FeatureCloud

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

In FeatureCloud, 1 participating client also takes the aggregating role and is called the coordinator. The app is implemented as a state machine, meaning that the app switches between states to perform different tasks. All states and their transitions are shown in Multimedia Appendix 1. After reading the local data and config files, minimizing the objective function using a federated Newton conjugate gradient is performed iteratively. Therefore, the local gradient and Hessian matrices are calculated and sent to the coordinator. The coordinator aggregates these data to obtain the global matrices, updates the weight vector \( \omega \), and broadcasts it to all clients. This is repeated until convergence.

A considerable advantage of the FeatureCloud platform is its native support of 2 very popular PETs, such as secure multiparty computation (SMPC). For applying SMPC, FeatureCloud supports a secure aggregation scheme for hiding locally exchanged parameters using additive secret sharing [26]. Through this, the exchanged local models are protected, and...
only the global aggregations are visible to attackers, clients, and the global aggregator. This is achieved by splitting the value that needs to be exchanged with the global aggregator into \( n \) shards, where \( n \) is the number of participating clients, and the sum of these \( n \) shards would result in the actual value \([23]\). Each shard is encrypted using a public key of each participant. These encrypted shards are shared with the global aggregator, sending them to the corresponding client holding the private key. The clients decrypt the received shards, sum them up, and send them back to the global aggregator, which sums up all received sums. This final sum results in the actual, nonhidden, global aggregate.

**Ethical Considerations**

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

**Evaluation**

We evaluated our approach using the developed FeatureCloud app on 3 benchmark data sets, all available via the scikit-survival package. The breast cancer data set (BRCA) \([27]\) contains the gene expression profiling of microarray experiments from 198 primary breast tumors, originally used to validate a 76-gene prognostic signature able to predict distant metastases in lymph node–negative patients with breast cancer. The German Breast Cancer Study Group 2 data set (GBSG2) \([28]\) contains data from a multicenter randomized clinical trial to compare the effectiveness of 3 versus 6 cycles of cyclophosphamide, methotrexate, and fluorouracil on recurrence-free and overall survival of 686 women. The observed parameters were hormonal therapy (yes or no), age of the patients, menopausal status (pre vs post), tumor size (in mm), tumor grade, number of positive tumor nodes, progesterone receptor (in fmol), and estrogen, as well as the censoring indicator and recurrence-free survival time (in days). The Worcester Heart Attack Study data set (WHAS500) \([29]\) contains data from 500 patients with acute myocardial infarction, collected during thirteen 1-year periods. Parameters were age, gender, initial heart rate, initial systolic and diastolic blood pressure, body mass index, history of cardiovascular disease, atrial fibrillation, cardiogenic shock, congestive heart complications, complete heart block, myocardial infarction order and type, vital status, and total length of follow-up.

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

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

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

\(^{a}\)BRCA: breast cancer data set.

\(^{b}\)GBSG2: German Breast Cancer Study Group 2 data set.

\(^{c}\)WHAS500: Worcester Heart Attack Study data set.

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

To evaluate the accuracy of our model, we used the Harrell concordance index, which was developed as a generalization of the area under the receiver operating characteristic curve for time-to-event models \([33]\). It corresponds to the probability of concordance between observed and predicted survival based on each pair of individuals. A c-index of 0.5 means that the model performs as well as a random guess, and a c-index of 1.0 means that the model predicts perfectly well.

After preprocessing, we performed a 3 × 3-fold cross-validation (CV) for a FeatureCloud workflow consisting of a federated normalization, the federated survival SVM, and a federated survival evaluation (c-index). We then compared our results
with the centralized analysis of every client and the merged data set (simulating a central data collection). Centralized analysis was performed using scikit-survival’s FastSurvivalSVM with a rank ratio of 0, α of 0.0001, true fit intercept, and a maximum of 50 iterations. The same hyperparameters were used for the federated analysis, respectively.

Privacy

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

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

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

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

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

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

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

Results

Performance

Our workflow delivered a highly similar model performance and model parameters for all federated analyses compared with the ones performed on the corresponding centralized data sets. The resulting c-indices to estimate the performance of our time-to-event models are depicted in Figure 2 [33]. For each data set (subplot), we show a boxplot consisting of the evaluated c-index for each CV split of our federated workflow with secure aggregation (green), federated workflow without secure aggregation (orange), and centralized calculation for each individual client (blue). The CV results show that our federated as well as the federated and secure aggregation approach perform highly similar to the centralized estimates. The calculation of the federated c-index in FeatureCloud causes small deviations in the c-index between centralized and federated. This is because FeatureCloud calculates a local c-index and aggregates to the mean c-indices of all sites. Therefore, it does not lead to the same c-index as a central computation would. The mean c-indices for the 4 data sets are in the range between 0.658 (GBSG2) and 0.76 (WHAS500). In contrast to the accuracy, achieving very high c-indices is rather difficult and depends very much on the problem. In a bioinformatics context, the lowest c-index of 0.658 (GBSG2) can be considered as moderate. The model achieves discrimination between individuals with different survival outcomes. However, it might not be of clinical utility and needs further refinement. The c-index of 0.76 (WHAS500) on the other hand, can be considered as good and has predictive value. Improving the predictive value of the models and increasing c-index was out of the scope of this work. A complete table of the results is available in Multimedia Appendix 2.
The model weights are nearly identical, with a maximum difference of only 0.001 and a mean difference of 0.0002 (Multimedia Appendices 1 and 3). These tiny differences between the weights of the central model and our model are negligible, as they do not change the overall prediction results and still lead to equal c-indices. The resulting model is therefore almost identical to the one that was trained on central data. A useful property of the linear survival SVM is, that the model weights can be used as a feature importance measure, which is also supported in our approach.

Besides calculating the feature importance from model weights directly, our federated survival SVM app uses Shapley additive explanations (SHAP), an explainable artificial intelligence framework for the interpretation of ML models [37]. Using SHAP, we compared the final models of the central, federated without secure aggregation, and federated with secure aggregation runs. For each data set, the SHAP shows highly similar model interpretations with a mean Pearson correlation of 0.991 between the central and the federated model without secure aggregation, and a mean Pearson correlation of 0.985 between the central model and the federated model with secure aggregation. A slightly worse correlation in the secure aggregation model is expected, as the masking of local parameters leads to floating-point issues. The worst correlation is shown in the microbiome data set (0.964), which can be explained by the high correlation between features in this data set. The results of the SHAP correlation analysis are listed in Multimedia Appendix 4 and the corresponding SHAP beeswarm plots are available in Multimedia Appendix 5.

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

**Runtime**

As shown in Figure 3, the runtime largely depends on the data set. In the case of FL, the number of iterations and, therefore, the number of data exchanges are the bottleneck. While the federated-only approach has linear runtime, the runtime of federated and secure aggregation is much worse and increases with an increasing number of clients. As described in the FeatureCloud publication, providing better privacy by hiding...
the exchanged parameters from the global aggregator, the simple additive secret sharing grows quadratic with the number of participants. Especially when many iterations and data exchanges are needed, this has a bad influence on the runtime of the FL implementation.

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

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

**FeatureCloud App**

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

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

**Discussion**

**Principal Findings**

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

**Comparison to Existing Work**

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

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

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

**Limitations**

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

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

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

**Conclusions**

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

**Acknowledgments**

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

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

Conflicts of Interest

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

Multimedia Appendix 1
State workflow of the survival support vector machine (SVM) FeatureCloud app and difference between coefficients.
[DOCX File, 244 KB - ai_v3i1e47652_app1.docx ]

Multimedia Appendix 2
C-indices of central, federated, and federated + secure aggregation analyses.
[XLSX File (Microsoft Excel File), 32 KB - ai_v3i1e47652_app2.xlsx ]

Multimedia Appendix 3
Coefficients of the trained survival support vector machines (SVMs).
[XLSX File (Microsoft Excel File), 243 KB - ai_v3i1e47652_app3.xlsx ]

Multimedia Appendix 4
Correlation of Shapley additive explanations (SHAP) values between central, federated, and federated + secure aggregation model.
[XLSX File (Microsoft Excel File), 10 KB - ai_v3i1e47652_app4.xlsx ]

Multimedia Appendix 5
Shapley additive explanations (SHAP) beeswarm plots for the different models.
[ZIP File (Zip Archive), 25020 KB - ai_v3i1e47652_app5.zip ]

Multimedia Appendix 6
Runtime of the federated survival support vector machine (SVM) training with 1, 3, and 5 clients.
[XLSX File (Microsoft Excel File), 11 KB - ai_v3i1e47652_app6.xlsx ]

Multimedia Appendix 7
Runtime of the federated survival support vector machine (SVM) with 1, 3, and 5 clients of a large sample size synthetic data set.
[XLSX File (Microsoft Excel File), 10 KB - ai_v3i1e47652_app7.xlsx ]

References


ML: machine learning
MSP: Metagenomic Species Pangenome
PET: privacy-enhancing technique
RSA: Rivest–Shamir–Adleman
SHAP: Shapley additive explanations
SMPC: secure multiparty computation
SVM: support vector machine
WHASS500: Worcester Heart Attack Study data set
Predictive Performance of Machine Learning–Based Models for Poststroke Clinical Outcomes in Comparison With Conventional Prognostic Scores: Multicenter, Hospital-Based Observational Study

Fumi Irie\textsuperscript{1,2*}, MD, PhD; Koutarou Matsumoto\textsuperscript{3*}, MPH, PhD; Ryu Matsuo\textsuperscript{1,2}, MD, PhD; Yasunobu Nohara\textsuperscript{4}, PhD; Yoshinobu Wakisaka\textsuperscript{2}, MD, PhD; Tetsuro Ago\textsuperscript{2,5}, MD, PhD; Naoki Nakashima\textsuperscript{6}, MD, PhD; Takanari Kitazono\textsuperscript{2,5}, MD, PhD; Masahiro Kamouchi\textsuperscript{1,5}, MD, PhD\textsuperscript{‡}

\textsuperscript{1}Department of Health Care Administration and Management, Graduate School of Medical Sciences, Kyushu University, Fukuoka, Japan
\textsuperscript{2}Department of Medicine and Clinical Science, Graduate School of Medical Sciences, Kyushu University, Fukuoka, Japan
\textsuperscript{3}Biostatistics Center, Graduate School of Medicine, Kurume University, Kurume, Japan
\textsuperscript{4}Big Data Science and Technology, Faculty of Advanced Science and Technology, Kumamoto University, Kumamoto, Japan
\textsuperscript{5}Center for Cohort Studies, Graduate School of Medical Sciences, Kyushu University, Fukuoka, Japan
\textsuperscript{6}Medical Information Center, Kyushu University Hospital, Fukuoka, Japan
\textsuperscript{‡}Fukuoka Stroke Registry Investigators
\textsuperscript{*}these authors contributed equally

\textbf{Corresponding Author:}
Masahiro Kamouchi, MD, PhD
Department of Health Care Administration and Management
Graduate School of Medical Sciences
Kyushu University
3-1-1 Maidashi
Higashi-ku
Fukuoka, 812-8582
Japan
Phone: 81 92 642 6960
Email: kamouchi.masahiro.736@m.kyushu-u.ac.jp

\textbf{Abstract}

\textbf{Background:} Although machine learning is a promising tool for making prognoses, the performance of machine learning in predicting outcomes after stroke remains to be examined.

\textbf{Objective:} This study aims to examine how much data-driven models with machine learning improve predictive performance for poststroke outcomes compared with conventional stroke prognostic scores and to elucidate how explanatory variables in machine learning–based models differ from the items of the stroke prognostic scores.

\textbf{Methods:} We used data from 10,513 patients who were registered in a multicenter prospective stroke registry in Japan between 2007 and 2017. The outcomes were poor functional outcome (modified Rankin Scale score >2) and death at 3 months after stroke. Machine learning–based models were developed using all variables with regularization methods, random forests, or boosted trees. We selected 3 stroke prognostic scores, namely, ASTRAL (Acute Stroke Registry and Analysis of Lausanne), PLAN (preadmission comorbidities, level of consciousness, age, neurologic deficit), and iScore (Ischemic Stroke Predictive Risk Score) for comparison. Item-based regression models were developed using the items of these 3 scores. The model performance was assessed in terms of discrimination and calibration. To compare the predictive performance of the data-driven model with that of the item-based model, we performed internal validation after random splits of identical populations into 80% of patients as a training set and 20% of patients as a test set; the models were developed in the training set and were validated in the test set. We evaluated the contribution of each variable to the models and compared the predictors used in the machine learning–based models with the items of the stroke prognostic scores.
**Results:** The mean age of the study patients was 73.0 (SD 12.5) years, and 59.1% (6209/10,513) of them were men. The area under the receiver operating characteristic curves and the area under the precision-recall curves for predicting poststroke outcomes were higher for machine learning–based models than for item-based models in identical populations after random splits. Machine learning–based models also performed better than item-based models in terms of the Brier score. Machine learning–based models used different explanatory variables, such as laboratory data, from the items of the conventional stroke prognostic scores. Including these data in the machine learning–based models as explanatory variables improved performance in predicting outcomes after stroke, especially poststroke death.

**Conclusions:** Machine learning–based models performed better in predicting poststroke outcomes than regression models using the items of conventional stroke prognostic scores, although they required additional variables, such as laboratory data, to attain improved performance. Further studies are warranted to validate the usefulness of machine learning in clinical settings.

(***JMIR AI** 2024;3:e46840) doi: [10.2196/46840](https://doi.org/10.2196/46840)

**KEYWORDS**
brain infarction; outcome; prediction; machine learning; prognostic score

**Introduction**

**Background**

Despite receiving the best available treatment, patients who have had a stroke may still experience disability or, in some cases, even face the risk of death [1,2]. Stroke clinicians try to predict patients’ outcomes as accurately as possible because accurate prognoses are a prerequisite for therapeutic decisions. Various stroke prognostic scores have been developed to support clinicians in predicting poststroke outcomes [3-8]. Nevertheless, prognostic scores have some disadvantages: generally, they limit the number of variables for ease of use at the bedside, and their validity needs to be reappraised over time, as the scoring criteria may become outdated with rapid progress in stroke care [9].

Meanwhile, recent advances in information technology have enabled the collection of a large amount of health information on individual patients [10,11]. Machine learning is considered a promising tool for improving the prediction accuracy of clinical outcomes for individual patients with stroke because of the ability of machine learning to deal with large and complex data [12-24].

However, several papers questioning the incremental value of machine learning have recently been published [25-27]. One study reported that machine learning algorithms did not perform better than traditional regression models for making prognoses in traumatic brain injury and recommended replicating studies in fields other than traumatic brain injury to ensure the generalizability of the findings [26]. Hitherto, few studies have directly compared the performance of data-driven models developed using machine learning methods and regression models based on conventional stroke prognostic scores in the field of outcome prediction after ischemic stroke [19,20,23]. In addition, calibration has not been adequately addressed in previous studies, and model performance has primarily been evaluated based on its discriminative ability [18-20].

**Objectives**

In this study, we aimed to examine whether machine learning can improve the predictive performance for poststroke outcomes beyond preexisting stroke prognostic scores. We also sought to elucidate the pattern of variables selected by machine learning algorithms to predict poststroke clinical outcomes. To this end, we analyzed the data of patients with acute ischemic stroke enrolled in a multicenter, hospital-based, prospective registry of stroke in Japan. We used 3 stroke prognostic scores, namely, Acute Stroke Registry and Analysis of Lausanne (ASTRAL) score [6], preadmission comorbidities, level of consciousness, age, and neurologic deficit (PLAN) score [7], and Ischemic Stroke Predictive Risk Score (iScore) [4,5], to create item-based regression models. We then compared the predictive performance of data-driven models developed using machine learning algorithms with that of item-based models in identical study populations. We also examined the explanatory variables used in data-driven models and compared them with the items of the conventional prognostic scores.

**Methods**

**Ethical Considerations**

The study protocol was approved by the institutional review boards of all hospitals (Kyushu University Institutional Review Board for Clinical Research: 22086-01; Kyushu Medical Center Institutional Review Board: R06-03; Clinical Research Review Board of Fukuokahigashi Medical Center: 29-C-38; Fukuoka Red Cross Hospital Institutional Review Board: 629; St Mary’s Hospital Research Ethics Review Committee: S13-0110; Steel Memorial Yawata Hospital Ethics Committee: 06-04-13; and Fukuoka Rosai Hospital Institutional Review Board: 21-8). Written informed consent was obtained from all patients or their family members.

**Data Source**

We used data from the Fukuoka Stroke Registry (FSR), a multicenter, hospital-based, prospective registry of patients with acute stroke. FSR enrolled patients with stroke hospitalized in 7 participating hospitals in Fukuoka, Japan, within 7 days of onset (University Hospital Medical Information Network Clinical Trial Registry: UMIN000000800). Details of the registry have been previously published [28,29]. In FSR, clinical data during routine stroke care in the hospitals were recorded along with baseline information on variables such as demographics, prior history, comorbidity, and functional level.
before stroke onset. The definitions of these variables have been previously described [28,29].

**Stroke Prognostic Scores**

The conventional stroke prognostic scores were used for comparison against data-driven prediction models. In this study, we selected prognostic scores based on the following criteria: they are multiitem and point-based scores using demographic and clinical information, they were developed to predict short-term outcomes after ischemic stroke, and they were externally validated. Consequently, 3 stroke prognostic scores, the ASTRAL score [6], PLAN score [7], and iScore [4,5], were used for comparative analysis. Items of these preexisting stroke prognostic scores were used as explanatory variables in item-based models (Multimedia Appendix 1).

**Study Populations**

FSR included 10,700 consecutive patients with acute ischemic stroke who were registered between June 2007 and May 2017. Ischemic stroke was diagnosed based on the sudden onset of a nonconvulsive and focal neurological deficit confirmed by brain imaging through computed tomography, magnetic resonance imaging, or both conducted upon admission. Of the 10,700 patients, 187 (1.7%) were lost to follow-up, and the remaining 10,513 (98.3%) were analyzed for 3 months post stroke.

Study patients were selected according to the inclusion and exclusion criteria of preexisting stroke prognostic scores to make the study populations identical between the item-based and machine learning–based models (Multimedia Appendix 2). Furthermore, we limited the study to patients with complete data, ensuring there were no missing variables across all data points. This approach aimed to prevent further reduction in the number of analyzed patients owing to list-wise deletion in regression models. The frequency of missing data is shown in Multimedia Appendix 3. Consequently, population 1, population 2, and population 3 were included in the analysis for comparison with the ASTRAL score, PLAN score, and iScore, respectively. Figure 1 illustrates the patient selection in each population.

**Study Outcomes**

The study outcomes were poor functional outcome and death at 3 months after stroke. Poor functional outcome was defined as a modified Rankin Scale score >2 at 3 months after stroke onset [30]. Death was defined as death from any cause within 3 months after stroke [30]. Interviewers on clinical outcomes were blinded to the patients’ backgrounds.

**Development of Predictive Models**

We performed logistic regression analysis to develop item-based models using the predictors of the ASTRAL score, PLAN score,
and iScore as explanatory variables (Multimedia Appendix 1). The predictors used in these models included age, time delay from onset to admission, stroke scale score, decreased level of consciousness, visual field defect, and abnormal glucose levels for the ASTRAL score; age, atrial fibrillation, congestive heart failure, cancer, preadmission dependence, decreased level of consciousness, leg weakness, arm weakness, and aphasia or neglect for the PLAN score; age, male sex, atrial fibrillation, congestive heart failure, renal dialysis, cancer, preadmission dependence, Canadian Neurological Scale score, stroke subtype, and abnormal glucose levels for the iScore. The categorization of predictors in the stroke prognostic scores was the same as that used in the original study for each score.

We used regularization methods (ridge regression [RR] and least absolute shrinkage and selection operator [LASSO] regression models) and ensemble decision tree models (random forest [RF] and Extreme Gradient Boosting [XGBoost]) for data-driven models based on machine learning algorithms [31-34]. All available variables were included in the development of data-driven models (Multimedia Appendix 3). The details of the model development are presented in Multimedia Appendix 4.

**Metrics of Model Performance**

The discriminative ability of each model was evaluated using the area under the receiver operating characteristic curve (AUROC) and the area under the precision-recall curve (AUPRC). AUPRC was calculated because it is a useful performance metric for unbalanced data of infrequent outcome events, such as death [35].

The calibration of each model was assessed using a calibration plot. Calibration plots were obtained by plotting the predicted and observed probabilities of the clinical outcomes in the 10 risk groups estimated using each predictive model. The Brier score was also used to assess the overall performance. The Brier score is defined as \(1/N \sum_{i=1}^{N} (pi-a_i)^2\), \(0 \leq \text{BS} \leq 1\), where \(pi\) is the predicted probability of the occurrence of an event ranging from 0 to 1, \(a_i\) indicates the event with binary outcomes (1 for observed or 0 for not observed), and \(N\) is the number of samples.

**Validation and Comparison of Models**

We performed internal validation of item-based and data-driven models after 100 repeated random splits into 80% of the patients as a training set and 20% of patients as a test set (Figure 2). The parameters in the training set were optimally tuned via 10-fold cross-validation in the data-driven models. After 100 random splits, the predictive models were developed by logistic regression using the items of the stroke prognostic scores (item-based model) and by machine learning using all variables (data-driven model) in the training set. The developed item-based and data-driven models were validated in the test set. The data sets for both training and testing were identical for the item-based and data-driven models. The median and 95% CI of the performance metrics, that is, AUROC, AUPRC, and Brier score, were calculated for each model using the results of the 100 repeated random splits. To directly compare the performance of the item-based and data-driven models (RR, LASSO, RF, and XGBoost), we compared the AUROC, AUPRC, and Brier score of the data-driven models with those of the corresponding item-based model. We repeated the comparison 100 times and calculated the times that the AUROC, AUPRC, and Brier score of data-driven models were better than those of the corresponding item-based model among the 100 repetitions.
Figure 2. Schematic diagram of the development and validation of the predictive models. All patients were randomly split into 80% of the development cohort as training data and 20% of the validation cohort as test data, which was repeated 100 times. Among the data-driven models, predictive models were developed based on ridge regression (RR), least absolute shrinkage and selection operator regression (LASSO), random forest (RF), and Extreme Gradient Boosting (XGBoost) using all available data after hyperparameter tuning in the development cohort. Logistic regression was used with predictors of stroke prognostic scores in the item-based models. The predictive models were validated using the test data of the validation cohort. In each split, the training and test data were identical between the data-driven and item-based models. ASTRAL: Acute Stroke Registry and Analysis of Lausanne; PLAN: preadmission comorbidities, level of consciousness, age, and neurologic deficit.

Evaluation of the Contribution of Variables
We evaluated the importance of the variables used in the item-based and data-driven models. To assess the contribution of each predictor to the item-based regression model, we calculated the rate of times when the association between each variable and clinical outcomes was statistically significant ($P<.05$) after 100 random splits. In the machine learning models, the magnitude of variable importance was evaluated in identical populations after 100 random splits (Multimedia Appendix 4).

We calculated the AUROC of the XGBoost model using various types of variables to assess how the addition of explanatory variables improves the predictive performance of the data-driven model. First, we constructed a model with age, sex, National Institutes of Health Stroke Scale (NIHSS) score, and preadmission modified Rankin Scale score (model 1). Then, 5 models were developed by adding items relating to preadmission status to model 1 (model 2), items relating to clinical data on admission to model 2 (model 3), items relating to brain imaging data to model 3 (model 4), and items relating to laboratory data to model 4 (model 5).

Statistical Analysis
We used the chi-square test, 2-tailed Student $t$ test, or Mann-Whitney $U$ test to compare the differences in baseline characteristics and clinical data, as appropriate [36]. Two-sided $P$ values <.05 were considered statistically significant.

All statistical analyses were performed using the R statistical package (R Development Core Team). This study was conducted in accordance with the Transparent Reporting of a Multivariable Prediction Model for Individual Prognosis or Diagnosis (TRIPOD) initiative [37].

Results
Baseline Variables and Clinical Outcomes
The mean age of the 10,513 patients was 73.0 (SD 12.5) years, and 59.1% (6209/10,513) of the patients were men. At 3 months after stroke, a poor functional outcome was found in 1204 (31.4%) of 3832 patients in population 1, 2209 (35.9%) of 6154 patients in population 2, and 2540 (37.1%) of 6855 patients in population 3. Within 3 months after stroke onset, 3% (113/3832), 3.6% (219/6154), and 3.7% (255/6855) of the patients died in population 1, population 2, and population 3, respectively.

First, we investigated the differences in the predictors of preexisting point-based stroke prognostic scores among patients according to poststroke clinical outcomes. Consequently, almost all variables significantly ($P<.05$) differed depending on the 3-month functional outcome (Table 1) and 3-month survival status (Multimedia Appendix 5) in addition to the predictors used in preexisting prognostic scores.
Table 1. Baseline data according to functional outcome at 3 months.

<table>
<thead>
<tr>
<th>Demographics</th>
<th>Overall (n=10,513)</th>
<th>mRS&lt;sup&gt;a&lt;/sup&gt; 0-2 (n=6405)</th>
<th>mRS 3-6 (n=4108)</th>
<th>P value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Age (y), mean (SD)</td>
<td>73.0 (12.5)</td>
<td>68.9 (12.0)</td>
<td>79.4 (10.4)</td>
<td>&lt;.001</td>
</tr>
<tr>
<td>Men, n (%)</td>
<td>6209 (59.1)</td>
<td>4257 (66.5)</td>
<td>1952 (47.5)</td>
<td>&lt;.001</td>
</tr>
<tr>
<td>Risk factors, n (%)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Hypertension</td>
<td>8485 (80.7)</td>
<td>5138 (80.2)</td>
<td>3347 (81.5)</td>
<td>.11</td>
</tr>
<tr>
<td>Diabetes mellitus</td>
<td>3607 (34.3)</td>
<td>2236 (34.9)</td>
<td>1371 (33.4)</td>
<td>.11</td>
</tr>
<tr>
<td>Atrial fibrillation</td>
<td>2743 (26.1)</td>
<td>1173 (18.3)</td>
<td>1570 (38.3)</td>
<td>&lt;.001</td>
</tr>
<tr>
<td>Smoking</td>
<td>2261 (23.1)</td>
<td>1717 (28.9)</td>
<td>544 (14.2)</td>
<td>&lt;.001</td>
</tr>
<tr>
<td>Comorbid conditions, n (%)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Congestive heart failure</td>
<td>919 (8.7)</td>
<td>423 (6.6)</td>
<td>496 (12.1)</td>
<td>&lt;.001</td>
</tr>
<tr>
<td>Kidney disease on dialysis</td>
<td>332 (3.2)</td>
<td>171 (2.7)</td>
<td>161 (3.9)</td>
<td>&lt;.001</td>
</tr>
<tr>
<td>Cancer</td>
<td>1552 (14.8)</td>
<td>774 (12.1)</td>
<td>778 (18.9)</td>
<td>&lt;.001</td>
</tr>
<tr>
<td>Previous history, n (%)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Previous myocardial infarction</td>
<td>505 (5.3)</td>
<td>242 (4.3)</td>
<td>263 (6.9)</td>
<td>&lt;.001</td>
</tr>
<tr>
<td>Preadmission functional status</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Preadmission mRS, median (IQR)</td>
<td>0 (0-1)</td>
<td>0 (0-0)</td>
<td>1 (0-3)</td>
<td>&lt;.001</td>
</tr>
<tr>
<td>Preadmission dependence (mRS score &gt;1), n (%)</td>
<td>2366 (22.5)</td>
<td>364 (5.7)</td>
<td>2002 (48.7)</td>
<td>&lt;.001</td>
</tr>
<tr>
<td>Onset-to-admission time, n (%)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>≤1 h</td>
<td>943 (9)</td>
<td>490 (7.7)</td>
<td>453 (11)</td>
<td>&lt;.001</td>
</tr>
<tr>
<td>≤3 h</td>
<td>1469 (14)</td>
<td>771 (12)</td>
<td>698 (17)</td>
<td></td>
</tr>
<tr>
<td>≤6 h</td>
<td>1141 (10.9)</td>
<td>644 (10.1)</td>
<td>497 (12.1)</td>
<td></td>
</tr>
<tr>
<td>≤24 h</td>
<td>3515 (33.4)</td>
<td>2090 (32.6)</td>
<td>1425 (34.7)</td>
<td></td>
</tr>
<tr>
<td>&gt;24 h</td>
<td>3445 (32.8)</td>
<td>2410 (37.6)</td>
<td>1035 (25.2)</td>
<td></td>
</tr>
<tr>
<td>Stroke subtype, n (%)</td>
<td></td>
<td></td>
<td></td>
<td>&lt;.001</td>
</tr>
<tr>
<td>Small vessel occlusion</td>
<td>2119 (20.2)</td>
<td>1724 (26.9)</td>
<td>395 (9.6)</td>
<td></td>
</tr>
<tr>
<td>Large artery atherosclerosis</td>
<td>1823 (17.3)</td>
<td>1006 (15.7)</td>
<td>817 (19.9)</td>
<td></td>
</tr>
<tr>
<td>Cardioembolism</td>
<td>2496 (23.7)</td>
<td>1054 (16.5)</td>
<td>1442 (35.1)</td>
<td></td>
</tr>
<tr>
<td>Other determined etiology</td>
<td>2146 (20.4)</td>
<td>1404 (21.9)</td>
<td>742 (18.1)</td>
<td></td>
</tr>
<tr>
<td>Undetermined</td>
<td>1929 (18.3)</td>
<td>1217 (19)</td>
<td>712 (17.3)</td>
<td></td>
</tr>
<tr>
<td>Neurological severity, median (IQR) or n (%)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>NIHSS&lt;sup&gt;b&lt;/sup&gt; score</td>
<td>3 (2-8)</td>
<td>2 (1-4)</td>
<td>8 (4-16)</td>
<td>&lt;.001</td>
</tr>
<tr>
<td>Severe stroke (NIHSS score &gt;10)</td>
<td>1938 (18.4)</td>
<td>291 (4.5)</td>
<td>1647 (40.1)</td>
<td>&lt;.001</td>
</tr>
<tr>
<td>Neurological deficits, n (%)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Decreased level of consciousness</td>
<td>3129 (30)</td>
<td>770 (12.1)</td>
<td>2359 (57.9)</td>
<td>&lt;.001</td>
</tr>
<tr>
<td>Leg weakness</td>
<td>5394 (51.9)</td>
<td>2357 (37.2)</td>
<td>3037 (75)</td>
<td>&lt;.001</td>
</tr>
<tr>
<td>Arm weakness</td>
<td>5634 (54.2)</td>
<td>2520 (39.7)</td>
<td>3114 (76.8)</td>
<td>&lt;.001</td>
</tr>
<tr>
<td>Aphasia or neglect</td>
<td>2912 (27.9)</td>
<td>946 (14.9)</td>
<td>1966 (48.3)</td>
<td>&lt;.001</td>
</tr>
<tr>
<td>Visual field defect</td>
<td>999 (9.6)</td>
<td>447 (7.0)</td>
<td>552 (13.6)</td>
<td>&lt;.001</td>
</tr>
<tr>
<td>Physiological data, mean (SD)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SBP&lt;sup&gt;c&lt;/sup&gt;, mm Hg</td>
<td>86.6 (18.2)</td>
<td>87.9 (17.8)</td>
<td>84.6 (18.6)</td>
<td>&lt;.001</td>
</tr>
<tr>
<td>DBP&lt;sup&gt;d&lt;/sup&gt;, mm Hg</td>
<td>159.8 (29.3)</td>
<td>160.4 (28.6)</td>
<td>158.8 (30.3)</td>
<td>.01</td>
</tr>
</tbody>
</table>

Note: mRS = modified Rankin Scale, SBP = systolic blood pressure, DBP = diastolic blood pressure.
<table>
<thead>
<tr>
<th></th>
<th>Overall (n=10,513)</th>
<th>mRS(^a) 0-2 (n=6405)</th>
<th>mRS 3-6 (n=4108)</th>
<th>(P) value</th>
</tr>
</thead>
<tbody>
<tr>
<td>BMI, kg/m(^2)</td>
<td>22.8 (3.8)</td>
<td>23.5 (3.6)</td>
<td>21.7 (3.9)</td>
<td>&lt;.001</td>
</tr>
<tr>
<td><strong>Laboratory data, median (IQR)</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Complete blood cell count</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>WBC(^c) (10(^3)/μL)</td>
<td>6.8 (5.6-8.4)</td>
<td>6.7 (5.5-8.2)</td>
<td>7.0 (5.7-8.9)</td>
<td>&lt;.001</td>
</tr>
<tr>
<td>RBC(^f) (10(^6)/μL)</td>
<td>436 (394-476)</td>
<td>449 (411-485)</td>
<td>416 (372-458)</td>
<td>&lt;.001</td>
</tr>
<tr>
<td>Hematocrit (%)</td>
<td>40.1 (36.5-44.3)</td>
<td>41.1 (37.9-44.0)</td>
<td>38.2 (34.6-41.9)</td>
<td>&lt;.001</td>
</tr>
<tr>
<td>Hemoglobin (g/dL)</td>
<td>13.5 (12.1-14.8)</td>
<td>14.0 (12.7-15.1)</td>
<td>12.8 (11.4-14.1)</td>
<td>&lt;.001</td>
</tr>
<tr>
<td>Platelet (10(^3)/μL)</td>
<td>20.2 (16.6-24.3)</td>
<td>20.6 (17.0-24.7)</td>
<td>19.5 (15.8-23.6)</td>
<td>&lt;.001</td>
</tr>
<tr>
<td><strong>Liver function</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>AST(^g) (U/L)</td>
<td>23 (19-29)</td>
<td>23 (19-29)</td>
<td>23 (19-30)</td>
<td>.001</td>
</tr>
<tr>
<td>ALT(^h) (U/L)</td>
<td>17 (12-24)</td>
<td>18 (13-25)</td>
<td>15 (11-22)</td>
<td>&lt;.001</td>
</tr>
<tr>
<td>LDH(^i) (U/L)</td>
<td>219 (186-266)</td>
<td>211 (181-254)</td>
<td>230 (195-285)</td>
<td>&lt;.001</td>
</tr>
<tr>
<td>ALP(^j) (U/L)</td>
<td>239 (195-295)</td>
<td>231 (190-284)</td>
<td>250 (203-312)</td>
<td>&lt;.001</td>
</tr>
<tr>
<td><strong>Kidney function</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>BUN(^k) (mg/dL)</td>
<td>16.0 (13.0-20.9)</td>
<td>15.3 (12.6-19.0)</td>
<td>17.9 (13.8-23.8)</td>
<td>&lt;.001</td>
</tr>
<tr>
<td>Creatinine (mg/dL)</td>
<td>0.8 (0.6-1.0)</td>
<td>0.8 (0.7-1.0)</td>
<td>0.8 (0.6-1.1)</td>
<td>&lt;.001</td>
</tr>
<tr>
<td>eGFR(^l) (mL/min/1.73 m(^2))</td>
<td>66.5 (51.2-81.5)</td>
<td>70.2 (55.9-83.8)</td>
<td>60.8 (44.8-76.5)</td>
<td>&lt;.001</td>
</tr>
<tr>
<td><strong>Glycemic control</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Glucose (mg/100 mL)</td>
<td>121 (103-156)</td>
<td>119 (103-154)</td>
<td>124 (105-158)</td>
<td>.001</td>
</tr>
<tr>
<td>Hemoglobin A(_{1c}) (%)</td>
<td>5.9 (5.6-6.6)</td>
<td>5.9 (5.6-6.6)</td>
<td>5.9 (5.5-6.5)</td>
<td>&lt;.001</td>
</tr>
<tr>
<td><strong>Inflammation</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>hsCRP(^m) (mg/dL)</td>
<td>1.5 (0.5-6.1)</td>
<td>1.0 (0.4-2.9)</td>
<td>3.9 (1.0-16.3)</td>
<td>&lt;.001</td>
</tr>
<tr>
<td><strong>Coagulation</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>PT-INR(^n)</td>
<td>1.0 (1.0-1.1)</td>
<td>1.0 (1.0-1.1)</td>
<td>1.1 (1.0-1.1)</td>
<td>&lt;.001</td>
</tr>
<tr>
<td>APTT(^o) (s)</td>
<td>29.7 (27.2-32.7)</td>
<td>29.5 (27.1-32.4)</td>
<td>30.1 (27.3-33.3)</td>
<td>&lt;.001</td>
</tr>
<tr>
<td>Fibrinogen (mg/dL)</td>
<td>304 (260-359)</td>
<td>297 (256-349)</td>
<td>315 (267-375)</td>
<td>&lt;.001</td>
</tr>
<tr>
<td>d-dimer (μg/mL)</td>
<td>0.9 (0.4-2.0)</td>
<td>0.6 (0.2-1.2)</td>
<td>1.7 (0.9-4.0)</td>
<td>&lt;.001</td>
</tr>
</tbody>
</table>

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

AUROCs varied depending on study populations, whereas differences between the machine learning algorithms were minimal in the same study population and for the same outcome. The AUROCs of data-driven models based on machine learning were generally higher than those of item-based models for predicting both 3-month poor functional outcome and all-cause death (Table 2). Similarly, AUPRCs were generally higher in data-driven models than in item-based models for predicting both poor functional outcome and all-cause death (Table 3). Regarding the Brier score, the data-driven models performed better than the item-based models (Table 4).

Table 2. Area under the receiver operating characteristic curve for predicting unfavorable clinical outcomes at 3 months using item-based and data-driven models.

<table>
<thead>
<tr>
<th>Poor functional outcome</th>
<th>Item-based model, median (95% CI)</th>
<th>Data-driven models, median (95% CI)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>RR^b</td>
<td>LASSO^c</td>
</tr>
<tr>
<td></td>
<td>RF^d</td>
<td>XGBoost^e</td>
</tr>
<tr>
<td>Population 1 (n=3832)</td>
<td>0.83 (0.80-0.85)</td>
<td>0.86 (0.83-0.89)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>0.86 (0.84-0.89)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>0.86 (0.84-0.88)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>0.86 (0.83-0.89)</td>
</tr>
<tr>
<td>Population 2 (n=6154)</td>
<td>0.88 (0.86-0.90)</td>
<td>0.91 (0.90-0.93)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>0.91 (0.90-0.93)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>0.91 (0.89-0.92)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>0.91 (0.89-0.93)</td>
</tr>
<tr>
<td>Population 3 (n=6855)</td>
<td>0.87 (0.85-0.89)</td>
<td>0.90 (0.89-0.92)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>0.90 (0.89-0.92)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>0.90 (0.89-0.92)</td>
</tr>
<tr>
<td>Death</td>
<td>0.77 (0.69-0.87)</td>
<td>0.87 (0.79-0.93)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>0.87 (0.78-0.92)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>0.89 (0.81-0.93)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>0.88 (0.82-0.93)</td>
</tr>
<tr>
<td>Population 2 (n=6154)</td>
<td>0.84 (0.80-0.89)</td>
<td>0.89 (0.85-0.92)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>0.88 (0.84-0.92)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>0.90 (0.86-0.93)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>0.90 (0.86-0.93)</td>
</tr>
<tr>
<td>Population 3 (n=6855)</td>
<td>0.82 (0.77-0.87)</td>
<td>0.88 (0.84-0.91)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>0.87 (0.83-0.90)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>0.89 (0.86-0.92)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>0.89 (0.85-0.91)</td>
</tr>
</tbody>
</table>

aThe study populations were selected according to the inclusion and exclusion criteria for the Acute Stroke Registry and Analysis of Lausanne (ASTRAL) score (population 1), the preadmission comorbidities, level of consciousness, age, and neurologic deficit (PLAN) score (population 2), and the Ischemic Stroke Predictive Risk Score (iScore; population 3).

bRR: ridge regression.

cLASSO: least absolute shrinkage and selection operator regression.

dRF: random forest.

eXGBoost: Extreme Gradient Boosting.

Table 3. Area under the precision-recall curve for predicting unfavorable clinical outcomes at 3 months using item-based and data-driven models.

<table>
<thead>
<tr>
<th>Poor functional outcome</th>
<th>Item-based model, median (95% CI)</th>
<th>Data-driven models, median (95% CI)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>RR^b</td>
<td>LASSO^c</td>
</tr>
<tr>
<td></td>
<td>RF^d</td>
<td>XGBoost^e</td>
</tr>
<tr>
<td>Population 1 (n=3832)</td>
<td>0.71 (0.66-0.75)</td>
<td>0.75 (0.71-0.79)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>0.75 (0.71-0.80)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>0.74 (0.69-0.79)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>0.75 (0.71-0.79)</td>
</tr>
<tr>
<td>Population 2 (n=6154)</td>
<td>0.83 (0.80-0.86)</td>
<td>0.87 (0.85-0.89)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>0.87 (0.85-0.90)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>0.87 (0.84-0.89)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>0.87 (0.85-0.89)</td>
</tr>
<tr>
<td>Population 3 (n=6855)</td>
<td>0.83 (0.80-0.85)</td>
<td>0.87 (0.85-0.89)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>0.87 (0.85-0.89)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>0.86 (0.84-0.88)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>0.87 (0.85-0.89)</td>
</tr>
<tr>
<td>Death</td>
<td>0.11 (0.06-0.24)</td>
<td>0.17 (0.08-0.32)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>0.17 (0.07-0.31)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>0.26 (0.13-0.44)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>0.24 (0.12-0.39)</td>
</tr>
<tr>
<td>Population 2 (n=6154)</td>
<td>0.17 (0.11-0.25)</td>
<td>0.27 (0.18-0.37)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>0.27 (0.18-0.38)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>0.29 (0.18-0.42)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>0.27 (0.16-0.35)</td>
</tr>
<tr>
<td>Population 3 (n=6855)</td>
<td>0.18 (0.11-0.25)</td>
<td>0.27 (0.16-0.36)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>0.27 (0.17-0.38)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>0.29 (0.19-0.42)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>0.28 (0.19-0.39)</td>
</tr>
</tbody>
</table>

aThe study populations were selected according to the inclusion and exclusion criteria for the Acute Stroke Registry and Analysis of Lausanne (ASTRAL) score (population 1), the preadmission comorbidities, level of consciousness, age, and neurologic deficit (PLAN) score (population 2), and the Ischemic Stroke Predictive Risk Score (iScore; population 3).

bRR: ridge regression.

cLASSO: least absolute shrinkage and selection operator regression.

dRF: random forest.

eXGBoost: Extreme Gradient Boosting.
Table 4. Brier score for predicting unfavorable clinical outcomes at 3 months using item-based and data-driven models\(^a\).

<table>
<thead>
<tr>
<th>Poor functional outcome</th>
<th>Item-based model, median (95% CI)</th>
<th>Data-driven models, median (95% CI)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>RR(^b)</td>
<td>LASSO(^c)</td>
</tr>
<tr>
<td>Population 1 (n=3832)</td>
<td>0.15 (0.14-0.17)</td>
<td>0.14 (0.12-0.15)</td>
</tr>
<tr>
<td>Population 2 (n=6154)</td>
<td>0.13 (0.12-0.14)</td>
<td>0.11 (0.10-0.12)</td>
</tr>
<tr>
<td>Population 3 (n=6855)</td>
<td>0.13 (0.12-0.15)</td>
<td>0.12 (0.11-0.13)</td>
</tr>
<tr>
<td>Death</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Population 1 (n=3832)</td>
<td>0.03 (0.02-0.03)</td>
<td>0.03 (0.02-0.03)</td>
</tr>
<tr>
<td>Population 2 (n=6154)</td>
<td>0.03 (0.02-0.04)</td>
<td>0.03 (0.02-0.04)</td>
</tr>
<tr>
<td>Population 3 (n=6855)</td>
<td>0.03 (0.02-0.04)</td>
<td>0.03 (0.02-0.04)</td>
</tr>
</tbody>
</table>

\(^a\)The study populations were selected according to the inclusion and exclusion criteria for the Acute Stroke Registry and Analysis of Lausanne (ASTRAL) score (population 1), the preadmission comorbidities, level of consciousness, age, and neurologic deficit (PLAN) score (population 2), and the Ischemic Stroke Predictive Risk Score (iScore; population 3).

\(^b\)RR: ridge regression.

\(^c\)LASSO: least absolute shrinkage and selection operator regression.

\(^d\)RF: random forest.

\(^e\)XGBoost: Extreme Gradient Boosting.

The predictive performance of data-driven models compared with the corresponding item-based model was examined by the frequency of the performance metrics (AUROC, AUPRC, and Brier score) of data-driven models, which were better than those of the corresponding item-based model in the identical training and test data sets after 100 repeated random splits (Table 5). Regarding poor functional outcome, the frequency exceeded 95% for all metrics in the data-driven models (RR, LASSO, RF, and XGBoost), indicating that the probability of the worse performance of data-driven models compared with the item-based model was <5%. Regarding death, the frequency was >95% for AUROC in all the data-driven models but did not always attain 95% for AUPRC or Brier score.

Calibration for predicting poor functional outcome was compared between the item-based and data-driven models (RR, LASSO, RF, and XGBoost) in population 1 for the ASTRAL score, in population 2 for the PLAN score, and in population 3 for the iScore. The prediction of poor functional outcome (Figure 3) and all-cause death (Figure 4) demonstrated concordance between the predicted and observed probabilities in the item-based models as well as in the data-driven models.
### Table 5. Predictive performance of data-driven models versus item-based models

<table>
<thead>
<tr>
<th></th>
<th>Poor functional outcome</th>
<th>Death</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>RR</td>
<td>LASSO</td>
</tr>
<tr>
<td>Population 1 (n=3832)</td>
<td>100</td>
<td>100</td>
</tr>
<tr>
<td>Population 2 (n=6154)</td>
<td>100</td>
<td>100</td>
</tr>
<tr>
<td>Population 3 (n=6855)</td>
<td>100</td>
<td>100</td>
</tr>
</tbody>
</table>

**AUROC**

<table>
<thead>
<tr>
<th></th>
<th>Poor functional outcome</th>
<th>Death</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>RR</td>
<td>LASSO</td>
</tr>
<tr>
<td>Population 1 (n=3832)</td>
<td>100</td>
<td>100</td>
</tr>
<tr>
<td>Population 2 (n=6154)</td>
<td>100</td>
<td>100</td>
</tr>
<tr>
<td>Population 3 (n=6855)</td>
<td>100</td>
<td>100</td>
</tr>
</tbody>
</table>

**AUPRC**

<table>
<thead>
<tr>
<th></th>
<th>Poor functional outcome</th>
<th>Death</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>RR</td>
<td>LASSO</td>
</tr>
<tr>
<td>Population 1 (n=3832)</td>
<td>100</td>
<td>100</td>
</tr>
<tr>
<td>Population 2 (n=6154)</td>
<td>100</td>
<td>100</td>
</tr>
<tr>
<td>Population 3 (n=6855)</td>
<td>100</td>
<td>100</td>
</tr>
</tbody>
</table>

*a* Data indicate the frequency that AUROC, AUPRC, and Brier score of data-driven models (RR, LASSO, RF, or XGBoost) exceeded those of item-based models in identical training and test sets after 100 repeated random splits.

*b* RR: ridge regression.

*c* LASSO: least absolute shrinkage and selection operator regression.

*d* RF: random forest.

*e* XGBoost: Extreme Gradient Boosting.

*f* AUROC: area under the receiver operating characteristic curve.

*g* AUPRC: area under the precision-recall curve.
Figure 3. Calibration of item-based and data-driven models for predicting poor functional outcome. Calibration for predicting poor functional outcome was compared between the item-based regression model and data-driven models (ridge regression [RR], least absolute shrinkage and selection operator regression [LASSO], random forest [RF], and Extreme Gradient Boosting [XGBoost]) in population 1 for the Acute Stroke Registry and Analysis of Lausanne (ASTRAL) score, population 2 for the preadmission comorbidities, level of consciousness, age, and neurologic deficit (PLAN) score, and population 3 for the Ischemic Stroke Predictive Risk Score (iScore). The patients were categorized into 10 groups stratified by the predicted probability of poor functional outcome in the test data. Observed probabilities (x-axis) were plotted against predicted probabilities (y-axis) in the 10 groups based on risk stratification. The results for the first 100 random splits are presented.
Figure 4. Calibration of item-based and data-driven models for predicting death. Calibration for predicting death was compared between the item-based regression model and data-driven models (ridge regression [RR], least absolute shrinkage and selection operator regression [LASSO], random forest [RF], and Extreme Gradient Boosting [XGBoost]) in population 1 for the Acute Stroke Registry and Analysis of Lausanne (ASTRAL) score, population 2 for the preadmission comorbidities, level of consciousness, age, and neurologic deficit (PLAN) score, and population 3 for the Ischemic Stroke Predictive Risk Score (iScore). The patients were categorized into 10 groups stratified by the predicted probability of death in the test data. Observed probabilities (x-axis) were plotted against predicted probabilities (y-axis) in the 10 groups based on risk stratification. The results for the first 100 random splits are presented.

Evaluation of Variables

Next, we evaluated how each variable contributed to the predictive performance of the item-based and data-driven models (RF and XGBoost) in population 1 (Figure 5), population 2 (Figure 6), and population 3 (Figure 7). The selected variables differed substantially between the study populations in the item-based models. Age, preadmission dependence, and neurological severity of stroke were important variables in predicting both poor functional outcome and death (Figures 5A, 7A; left panels). Age and neurological deficit signs (arm or leg weakness and loss of consciousness) were the most frequently used variables for predicting poor functional outcome (Figures 5A, 6A, and 7A; middle and right panels) in RF and XGBoost.

In contrast, variables not used in the item-based models, such as d-dimer, high-sensitivity C-reactive protein, fibrinogen, and BMI, were the most frequently used variables by RF and XGBoost (Figures 5B, 6B, and 7B; middle and right panels) in predicting death.

We also investigated how the addition of variables increased the predictive performance of XGBoost. As a result, the AUROC for poor functional outcome did not substantially increase even when explanatory variables other than key predictors were added to model 1 (Figure 8; open circles). Conversely, the AUROC for all-cause death linearly increased with the addition of other variables to the models, particularly items from laboratory data (Figure 8; closed circles).
Figure 5. Comparison of variable importance between items of the Acute Stroke Registry and Analysis of Lausanne (ASTRAL) score and explanatory variables in machine learning model in population 1. The contribution of each variable to the models in predicting poor functional outcome (A) and death (B) is shown. The patients were selected based on the ASTRAL criteria (population 1). In item-based regression models, the percentage indicates the rate of times when its association with clinical outcomes was statistically significant ($P < .05$). In machine learning models, the top 10 variables are shown according to the magnitude of variable importance. Boxes, vertical lines in the boxes, and horizontal bars indicate IQR, median, and minimal or maximal range, respectively. NIHSS: National Institutes of Health Stroke Scale, hsCRP: high-sensitivity C-reactive protein, LOC: loss of consciousness, mRS: modified Rankin Scale, BMI: body mass index, WBC: white blood cell count, LDH: lactate dehydrogenase, HbA1c: hemoglobin A1c, Fib: fibrinogen, Plt: platelet count, RBC: red blood cell count, ALP: alkaline phosphatase, Ht: hematocrit, Hb: hemoglobin, BUN: blood urea nitrogen, LDH: lactate dehydrogenase, PT-INR: international normalized ratio of prothrombin time.
Figure 6. Comparison of variable importance between items of the preadmission comorbidities, level of consciousness, age, and neurologic deficit (PLAN) score and explanatory variables in machine learning model in population 2. The contribution of each variable to the models in predicting poor functional outcome (A) and death (B) is shown. The patients were selected based on the PLAN score criteria (population 2). In item-based regression models, the percentage indicates the rate of times when its association with clinical outcomes was statistically significant ($P<.05$). In machine learning models, the top 10 variables are shown according to the magnitude of variable importance. Boxes, vertical lines in the boxes, and horizontal bars indicate IQR, median, and minimal or maximal range, respectively. mRS: modified Rankin Scale, LOC: loss of consciousness, hsCRP: high-sensitivity C-reactive protein, BMI: body mass index, Hb: hemoglobin, WBC: white blood cell count, Plt: platelet count, Fib: fibrinogen, RBC: red blood cell count, LDH: lactate dehydrogenase, Ht: hematocrit, ALP: alkaline phosphatase, PT-INR: international normalized ratio of prothrombin time.
Figure 7. Comparison of variable importance between items of Ischemic Stroke Predictive Risk Score (iScore) and explanatory variables in machine learning model in population 3. The contribution of each variable to the models in predicting poor functional outcome (A) and death (B) is shown. The patients were selected according to the iScore criteria (population 3). In item-based regression models, the percentage indicates the rate of times when its association with clinical outcomes was statistically significant ($P<.05$). In machine learning models, the top 10 variables are shown according to the magnitude of variable importance. Boxes, vertical lines in the boxes, and horizontal bars indicate IQR, median, and minimal or maximal range, respectively. NIHSS: National Institutes of Health Stroke Scale, CNS: Canadian Neurological Scale, mRS: modified Rankin Scale, LOC: loss of consciousness, hsCRP: high-sensitivity C-reactive protein, BMI: body mass index, Hb: hemoglobin, WBC: white blood cell count, Fibrinogen, RBC: red blood cell count, Plt: platelet count, Ht: hematocrit, LDH: lactate dehydrogenase, ALP: alkaline phosphatase, PT-INR: international normalized ratio of prothrombin time.

Figure 8. Improvement of discrimination in a data-driven model by adding different types of data. The area under the receiver operating characteristic curves (AUROCs) for predicting poor functional outcome (open circles) and death (closed circles) were compared among the 5 models, which used different types of variables. A data-driven model was developed for each population using Extreme Gradient Boosting. Vertical bars indicate the 95th percentile after 100 random splits. The variables used for the models were as follows: model 1: age, sex, National Institutes of Health Stroke Scale score, and preadmission modified Rankin Scale score; model 2: model 1 plus clinical data before admission (eg, risk factors, comorbid conditions, previous history, family history, and prestroke medication); model 3: model 2 plus clinical data on admission (eg, onset-to-admission time, ambulance use, BMI, and physiological data); model 4: model 3 plus brain imaging data (eg, site of lesion, side of lesion, and stroke subtype); and model 5: model 4 plus laboratory data.
Discussion

Principal Findings
This study, which analyzed comprehensive clinical data from a multicenter, hospital-based stroke registry, yielded the following major findings. The performance of item-based regression models using the predictors of 3 conventional stroke prognostic scores was fair in predicting clinical outcomes at 3 months after ischemic stroke in our cohort, despite differences in clinical and social backgrounds from the original cohorts of scores. Data-driven models based on machine learning algorithms exhibited better performance when compared with item-based models in identical study populations. The importance of variables in RF and XGBoost appeared to differ from that in item-based models when predicting death within 3 months. The addition of nonconventional factors, such as laboratory data, to the XGBoost model improved its predictive ability for 3-month mortality.

Predictive Performance of Models
Thus far, only a limited number of studies have evaluated the predictive performance of machine learning–based models compared with those of stroke prognostic scores [19,20,23]. All these studies were performed in single-center registries or under specific conditions, such as large vessel occlusion in ischemic stroke. Furthermore, previous studies mainly focused on AUROC for assessing predictive performance, although other metrics, such as measures of calibration, are necessary to fully evaluate the performance of models [38]. This study was conducted using a multicenter registry database and several performance metrics. Our study demonstrated that data-driven models developed using machine learning algorithms can perform reasonably well in predicting the 3-month clinical outcomes of patients with acute ischemic stroke. Generally, data-driven models performed better than conventional prognostic scores when both were compared in identical study populations.

This study also demonstrates that the model performance largely depends on the study populations. The study populations varied in terms of both size and patient characteristics, such as prestroke dependency, time from onset to admission, and use of thrombolytic therapy. The variability in AUROC, AUPRC, and Brier scores between the study populations was as large as that between the models. Moreover, the model performance varied depending on the outcomes to be predicted: AUPRCs were substantially decreased for the prediction of death, which is a less frequent event than the poor functional outcome. These findings underscore the reiterated importance of sample size, the number of outcome events, and data quality of the study cohorts where models are to be developed and validated [25,39,40].

Variables in Models
In this study, age, preadmission dependence, and variables related to neurological deficits were identified as important predictors for the prediction of poor functional outcome in both item-based regression models and data-driven models using RF and XGBoost. These are well-known risk factors for poor functional outcome and are also used for predicting death in stroke prognostic scores [4,5,7]. However, BMI and items related to laboratory data, such as D-dimer, high-sensitivity C-reactive protein, and fibrinogen, were found to be the most important variables for predicting death in RF and XGBoost. Indeed, the association between poststroke clinical outcomes and markers of inflammation and hypereooagulation has become a recent research topic [41,42]. Machine learning algorithms can be a promising tool to identify novel factors to be considered in making prognoses for stroke because they can maximize the use of data without arbitrary assumptions and procedures.

Clinical Implications
The ability of machine learning to derive a model that best fits the data on a given cohort is appealing for making prognoses. Prognostic scores with prespecified items may not fit all cohorts because heterogeneity must exist between study cohorts in race or ethnic groups, general health conditions, socioeconomic status, and health care systems. In addition, stroke prognostic scores are at risk of getting outdated over time, as advances in stroke care continuously improve clinical outcomes in patients with stroke [43,44]. However, our analysis suggests that the 3 conventional prognostic scores can perform sufficiently well in our cohort, despite the fact that the original studies that developed the scores had patients with different medical backgrounds and during different study periods. This finding demonstrates the robustness of outcome prediction using regression models in terms of generalizability. Furthermore, considering nonlinear and interaction effects might not be crucial for outcome prediction after ischemic stroke, as the simple regression models worked well in our study.

Point-based stroke prognostic scores are convenient and helpful for making prompt decisions at the bedside. Generally, prognostic scores comprise only a handful of variables on which information can be obtained easily. This advantage in the practicability of the prognostic scores is important in acute stroke care settings. Machine learning algorithms require more data than conventional prognostic scores to reach acceptable performance levels [39], and the data required by machine learning algorithms to realize better performance, such as laboratory data, may not always be available, although they can improve the predictive performance of models. Therefore, further studies are needed to fully assess the incremental value of machine learning–based models in daily clinical practice.

Strengths and Limitations
This study has several strengths. We assessed and compared the predictive accuracy of prognostic scores against data-driven models, using information from a multicenter, prospective registry of individuals diagnosed with acute stroke. We were able to use several variables, including laboratory data–related items, owing to the detailed clinical data available in the registry. Moreover, comparisons of models were made using various performance metrics. However, this study has also several limitations. First, the selection of patients may have led to bias, although the inclusion and exclusion criteria were identical to those reported in the original studies of the prognostic scores. Second, there were missing data for the baseline variables and clinical outcomes, which may have also led to selection bias.
Third, the possibility of overfitting cannot be completely ruled out, despite the predictive models constituted by the training set being fitted to the test set. Finally, this study included only patients with acute ischemic stroke who were hospitalized in tertiary care centers in a restricted region of Japan. Generalizability should be assessed in other settings and for other diseases.

Conclusions
This study suggests that data-driven models based on machine learning algorithms can improve predictive performance by using diverse types of variables, such as laboratory data–related items. The clinical outcomes of individual patients can be automatically estimated using machine learning algorithms if a large amount of data can be directly drawn from electronic health records. This possibility of making automated and personalized prognoses is an appealing property of data-driven prediction. However, the arrangement of an appropriate electronic infrastructure is indispensable for enabling data collection, and the development of such infrastructure requires time and cost. It is worth noting that conventional prognostic scores can achieve sufficient performance in making stroke prognoses with only a limited number of variables. In the near future, it seems feasible to explore the improvement of preexisting prognostic scores by incorporating novel predictors identified by machine learning algorithms, given the significant investment necessary to fully use machine learning.

Acknowledgments
This study was supported by the Japan Society for Promotion of Science KAKENHI (grants JP21H03165, JP21K19648, 21K10330, and JP22K10386) and the Ministry of Health, Labour and Welfare AC Program (grant JPMH21446713). The authors thank all the Fukuoka Stroke Registry investigators and their hospitals for participating in this study and all the clinical research coordinators from the Hisayama Research Institute for Lifestyle Diseases for their help in obtaining informed consent and collecting clinical data. Participating hospitals in the Fukuoka Stroke Registry included Kyushu University Hospital (Fukuoka, Japan), National Hospital Organization Kyushu Medical Center (Fukuoka, Japan), National Hospital Organization Fukuoka-Higashi Medical Center (Koga, Japan), Fukuoka Red Cross Hospital (Fukuoka, Japan), St Mary’s Hospital (Kurume, Japan), Steel Memorial Yawata Hospital (Kitakyushu, Japan), and Japan Labor Health and Welfare Organization Kyushu Rosai Hospital (Kitakyushu, Japan). Steering committee and research working group members of the Fukuoka Stroke Registry were Takao Ishitsuka, MD, PhD (Fukuoka Mirai Hospital, Fukuoka, Japan); Setsuro Ibayashi, MD, PhD (Chair, Seiai Rehabilitation Hospital, Onojo, Japan); Kenji Kusuda, MD, PhD (Seiai Rehabilitation Hospital, Onojo, Japan); Kenichiro Fujii, MD, PhD (Japan Seafarers Relief Association Moji Ekisaikai Hospital, Kitakyushu, Japan); Tetsuhiko Nagao, MD, PhD (Safety Monitoring Committee, Seiai Rehabilitation Hospital, Onojo, Japan); Yasushi Okada, MD, PhD (Vice-Chair, National Hospital Organization Kyushu Medical Center, Fukuoka, Japan); Masahiro Yasaka, MD, PhD (National Hospital Organization Kyushu Medical Center, Fukuoka, Japan); Hiroaki Ooboshi, MD, PhD (Fukuoka Dental College Medical and Dental Hospital, Fukuoka, Japan); Takanari Kitazono, MD, PhD (Principal Investigator, Kyushu University, Fukuoka, Japan); Katsumi Irie, MD, PhD (Hakujyuji Hospital, Fukuoka, Japan); Tsyuoshi Omae, MD, PhD (Imazu Red Cross Hospital, Fukuoka, Japan); Kazunori Toyoda, MD, PhD (National Cerebral and Cardiovascular Center, Suita, Japan); Hiroshi Nakane, MD, PhD (National Hospital Organization Fukuoka-Higashi Medical Center, Koga, Japan); Masahiro Kamouchi, MD, PhD (Kyushu University, Fukuoka, Japan); Hiroshi Sugimori, MD, PhD (National Hospital Organization Kyushu Medical Center, Fukuoka, Japan); Shuji Arakawa, MD, PhD (Steel Memorial Yawata Hospital, Kitakyushu, Japan); Kenji Fukuda, MD, PhD (St Mary’s Hospital, Kurume, Japan); Tetsuro Ago, MD, PhD (Kyushu University, Fukuoka, Japan); Jiro Kitayama, MD, PhD (Fukuoka Red Cross Hospital, Fukuoka, Japan); Shigeru Fujimoto, MD, PhD (Jichi Medical University, Shimotsuke, Japan); Shoji Arihiro, MD (Japan Labor Health and Welfare Organization Kyushu Rosai Hospital, Kitakyushu, Japan); Junya Kuroda, MD, PhD (National Hospital Organization Fukuoka-Higashi Medical Center, Koga, Japan); Yoshinobu Wakisaka, MD, PhD (Kyushu University Hospital, Fukuoka, Japan); Yoshihisa Fukushima, MD (St Mary’s Hospital, Kurume, Japan); Ryu Matsuo, MD, PhD (Secretariat, Kyushu University, Fukuoka, Japan); Fumi Irie, MD, PhD (Kyushu University, Fukuoka, Japan); Kuniiyuki Nakamura, MD, PhD (Kyushu University Hospital, Fukuoka, Japan); and Takuya Kiyohara, MD, PhD (Kyushu University Hospital, Fukuoka, Japan).

Conflicts of Interest
None declared.
Multimedia Appendix 3
Rates of missing values.

[DOCX File, 38 KB - ai_v3i1e46840_app3.docx]

Multimedia Appendix 4
R programs for the development of machine learning–based models.

[DOCX File, 33 KB - ai_v3i1e46840_app4.docx]

Multimedia Appendix 5
Baseline data according to death within 3 months.

[DOCX File, 40 KB - ai_v3i1e46840_app5.docx]

References


Abbreviations

ASTRAL: Acute Stroke Registry and Analysis of Lausanne
AUPRC: area under the precision-recall curve
AUROC: area under the receiver operating characteristic curve
FSR: Fukuoka Stroke Registry
iScore: Ischemic Stroke Predictive Risk Score
LASSO: least absolute shrinkage and selection operator
PLAN: preadmission comorbidities, level of consciousness, age, and neurologic deficit
RF: random forest
RR: ridge regression
TRIPOD: Transparent Reporting of a Multivariable Prediction Model for Individual Prognosis or Diagnosis
XGBoost: Extreme gradient boosting

©Fumi Irie, Koutarou Matsumoto, Ryu Matsuou, Yasunobu Nohara, Yoshinobu Wakisaka, Tetsuro Ago, Naoki Nakashima, Takanari Kitazono, Masahiro Kamouchi. Originally published in JMIR AI (https://ai.jmir.org), 11.01.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR AI, is properly cited. The complete bibliographic information, a link to the original publication on https://www.ai.jmir.org/, as well as this copyright and license information must be included.
Original Paper

Risk Perception, Acceptance, and Trust of Using AI in Gastroenterology Practice in the Asia-Pacific Region: Web-Based Survey Study

Wilson WB Goh, Kendrick YA Chia, Max FK Cheung, Kalya M Kee, BA, MSc; May O Lwin, BA, MBA, PhD; Peter J Schulz, BA, MA, PhD; Minhu Chen, MBBS, PhD; Kaichun Wu, MD, PhD; Simon SM Ng, MBChB, MD; Rashid Lui, MBChB; Tiing Leong Ang, MBBS, MRCP, UK; Khay Guan Yeoh, MBBS, MMed; Han-mo Chiu, MD, PhD; Deng-kyu-kyu, MD, PhD; Joseph JY Sung, MD, PhD

1Lee Kong Chian School of Medicine, Nanyang Technological University, Singapore, Singapore
2School of Biological Sciences, Nanyang Technological University, Singapore, Singapore
3Center for Biomedical Informatics, Nanyang Technological University, Singapore, Singapore
4Wee Kim Wee School of Communication and Information, Nanyang Technological University, Singapore, Singapore
5The First Affiliated Hospital, Sun Yat-sen University, Guangzhou, China
6Xijing Hospital, Fourth Military Medical University, Xi’an, China
7Department of Surgery, Faculty of Medicine, The Chinese University of Hong Kong, Hong Kong, China (Hong Kong)
8Prince of Wales Hospital, Hospital Authority, Hong Kong, China (Hong Kong)
9Department of Gastroenterology and Hepatology, Changi General Hospital, SingHealth, Singapore, Singapore
10Department of Gastroenterology and Hepatology, National University Hospital, National University Health System, Singapore, Singapore
11Department of Medicine, Yong Loo Lin School of Medicine, National University of Singapore, Singapore
12Department of Internal Medicine, National Taiwan University Hospital, Taiwan, China
13Department of Internal Medicine, College of Medicine, National Taiwan University, Taiwan, China
14Kaohsiung Medical University, Taiwan, China

Corresponding Author:
Wilson WB Goh, BSc, MSc, PhD
Lee Kong Chian School of Medicine
Nanyang Technological University
Singapore
Experimental Medicine Building, 59 Nanyang Drive
Singapore, 636921
Singapore
Phone: 65 67911744
Email: wilsongoh@ntu.edu.sg

Abstract

Background: The use of artificial intelligence (AI) can revolutionize health care, but this raises risk concerns. It is therefore crucial to understand how clinicians trust and accept AI technology. Gastroenterology, by its nature of being an image-based and intervention-heavy specialty, is an area where AI-assisted diagnosis and management can be applied extensively.

Objective: This study aimed to study how gastroenterologists or gastrointestinal surgeons accept and trust the use of AI in computer-aided detection (CADe), computer-aided characterization (CADx), and computer-aided intervention (CADi) of colorectal polyps in colonoscopy.

Methods: We conducted a web-based questionnaire from November 2022 to January 2023, involving 5 countries or areas in the Asia-Pacific region. The questionnaire included variables such as background and demography of users; intention to use AI, perceived risk; acceptance; and trust in AI-assisted detection, characterization, and intervention. We presented participants with 3 AI scenarios related to colonoscopy and the management of colorectal polyps. These scenarios reflect existing AI applications in colonoscopy, namely the detection of polyps (CADe), characterization of polyps (CADx), and AI-assisted polypectomy (CADi).
Results: In total, 165 gastroenterologists and gastrointestinal surgeons responded to a web-based survey using the structured questionnaire designed by experts in medical communications. Participants had a mean age of 44 (SD 9.65) years, were mostly male (n=116, 70.3%), and mostly worked in publicly funded hospitals (n=110, 66.67%). Participants reported relatively high exposure to AI, with 111 (67.27%) reporting having used AI for clinical diagnosis or treatment of digestive diseases. Gastroenterologists are highly interested to use AI in diagnosis but show different levels of reservations in risk prediction and acceptance of AI. Most participants (n=112, 72.72%) also expressed interest to use AI in their future practice. CADe was accepted by 83.03% (n=137) of respondents, CADx was accepted by 78.79% (n=130), and CADi was accepted by 72.12% (n=119). CADe and CADx were trusted by 85.45% (n=141) of respondents and CADi was trusted by 72.12% (n=119). There were no application-specific differences in risk perceptions, but more experienced clinicians gave lesser risk ratings.

Conclusions: Gastroenterologists reported overall high acceptance and trust levels of using AI-assisted colonoscopy in the management of colorectal polyps. However, this level of trust depends on the application scenario. Moreover, the relationship among risk perception, acceptance, and trust in using AI in gastroenterology practice is not straightforward.

JMIR AI 2024;3:e50525 doi:10.2196/50525

KEYWORDS
artificial intelligence; delivery of health care; gastroenterology; acceptance; trust; adoption; survey; surveys; questionnaire; questionnaires; detect; detection; colonoscopy; gastroenterologist; gastroenterologists; internal medicine; polyp; polyps; surgeon; surgeons; surgery; surgical; colorectal

Introduction

Artificial intelligence (AI) has made groundbreaking technological advancements in medical image interpretation [1]; diagnosis assistance; risk assessment for various conditions [2]; outcome prognostication [3]; and in certain areas, treatment suggestion [4] and partaking in surgical intervention [5]. Studies of AI trust and acceptance among clinicians are becoming increasingly important. This is because trust and acceptance of AI technology are seen as preconditions for clinical workflow integration [6]. Currently, trust has already been demonstrated by several studies as one of the main determinants in driving the adoption of AI in health care [7,8]. One study showed that within a general home-based health care setting—where AI is applied on the internet of things—based devices to monitor patients’ health—risk perception, acceptance, and trust are related concepts that govern the ultimate use of the developed technology [9]. A separate study [10] conducted on the use of an AI-based system in the application of a Blood Utilization Calculator showed that its trust and use were determined by perceived risk and expectancy (in our context, acceptance). It was demonstrated that high perceived risk reduced trust and subsequent use. While the clinical evidence of accuracy in the diagnosis and prognosis of AI is accumulating, the level of trust and acceptance by clinicians requires more attention [6]. We identified that gastroenterology, by its very nature of having heavy usage of image-based diagnosis (eg., computed tomography, magnetic resonance imaging, endoscopy, and histology) and surgical or endoscopic intervention, will be one of the specialties that may readily use AI technologies in clinical management [11,12]. Yet, there is little research on AI risk perception, acceptance, and trust among gastroenterologists.

To our knowledge, most published research surveys trust in a more general manner. One such recent example is the survey on gastrointestinal (GI) health care in 2022, which covered clinicians’ perspectives in a general way [13]. However, such surveys lack granularity. It is impossible to know under what circumstances do clinicians become less trusting or accepting or become more concerned about the deployments of AI. Moreover, there is a lack of explicit modeling from collected data to relate patterns of risk perception, acceptance, and trust among practitioners. There are existing models [14,15] that explore parts of the interactions among these 3 factors. However, because these explorations cover only partial relationships and interactions, we feel that these may be inadequate for modeling real-world dynamics. Therefore, having more comprehensive models would allow for a better understanding of the various factors underpinning how clinicians come to trust, accept, and eventually use AI. This knowledge would help in formulating successful implementation of AI tools in real-world environments.

In this study, we aim to understand the trust and acceptance among gastroenterologists, with a specific focus on the Asia-3Pacific region. We hypothesize is that risk perception, acceptance, and trust will change according to the scenario (computer-aided detection [CADe], computer-aided characterization [CADx], or computer-aided intervention [CADi]), with different levels of invasiveness. A blueprint of a survey that examines contextual responses toward screening colonoscopy with polypectomy in clinical environments is provided. Using our collected data, we attempt to elucidate how risk perception, acceptance, and trust interactions can be modeled and studied. These contributions collectively enhance our understanding of complex factors influencing the integration of AI in medical practice.

Methods

Survey

We used a structured questionnaire (Multimedia Appendix 1) to conduct a survey in English by inviting gastroenterologists or GI surgeons from the Asia-Pacific region through open invitations to various medical associations. The questionnaire was based on the expectancy-value framework, major constructs
of the Theory of Planned Behaviour research framework [16], and the Technology Acceptance Model measures [17]. Items in the questionnaire for testing risk perception, acceptance, and trust were adapted from various other studies [18,19], with some including items from validated constructs in questionnaires. These questions are then adapted into scenarios covering detection (CADe), characterization (CADx), or intervention (CADi), with different levels of invasiveness characterization and intervention for colonoscopic detection and polypectomy (see Textbox 1 for items used to evaluate these aspects).

Most items were rated on a 7-point Likert scale, where 7 denotes strong agreement. To assess risk perception, acceptance, and trust, we presented participants with 3 different AI applications related to colonoscopy and the management of colorectal polyps. These scenarios, reflecting existing AI applications in GI, involve the detection of polyps (CADe), characterization of the nature of polyps (CADx), and treatment procedures (CADi), respectively (see Table 1 and Textbox 1). Table 2 displays measurement items.

In this study, the three key elements for assessment are (1) risk perception, (2) acceptance, and (3) trust. Risk perception refers to an individual’s subjective assessment or understanding of the potential hazards, threats, or uncertainties associated with a particular situation or activity. It involves the process of evaluating and interpreting information about risk, considering factors such as the severity of potential consequences [20,21]. Acceptance is the mental and emotional state of acknowledging and accommodating a new concept or innovation into one’s beliefs, behaviors, or practices. Trust is defined as belief or confidence in the reliability, credibility, and integrity of a person, system, or technology leading to usage or action [20,21]. Acceptance may precede trust in the adoption of new technologies, but trust plays a crucial role in establishing a strong foundation for sustained usage and effective integration of AI into medical practice. Risk perception, acceptance, and trust may interact with each other and other factors stemming from professional, technological, and personal sources. The conceptual framework presented in Figure 1 illustrates the intricate interplay among sociodemographic variables, AI acceptance, trust, perceived risk, and outcomes [22]. Our study aims to contribute to this understanding not by testing individual relationships within this conceptual framework but by exploring how trust, risk, and acceptance are possibly interconnected in the context of AI-supported applications in gastroenterology.

**Computer-aided detection**
- Imagine you are attending an informal meeting of colleagues. Your colleagues are not experts in artificial intelligence and have about the same amount of understanding as you do. The conversation turns to innovation in medicine, especially machine learning algorithms and their potential to assist in the interpretation of medical imagery in the early detection of colon cancer. One of the colleagues speaks about a patient who underwent a colonoscopy which was assisted by a machine learning algorithm. When the algorithm indicated that the patient had a colonic polyp, the colleague asked for an additional biopsy. It turned out that the result produced by the algorithm was correct (use the following scale: 1=have major doubts to 4=neutral to 7=fully believe).

**Computer-aided characterization**
- The second colleague reported that the machine learning algorithm is also capable of correctly classifying whether the colonic polyp was adenomatous or hyperplastic (use the following scale: 1=have major doubts to 4=neutral to 7=fully believe).

**Computer-aided intervention**
- Now suppose a third colleague told you that a machine learning algorithm can be applied to guide interventions. Endoscopists need a targeted biopsy from specific locations that harbor the lesion. The third colleague said that the algorithm can guide a biopsy needle more precisely than a human, using ultrasound imaging (use the following scale: 1=have major doubts to 4=neutral to 7=fully believe).

**Table 1.** Scenarios demonstrating AI use in gastroenterology practice from detection to characterization and intervention.

<table>
<thead>
<tr>
<th>Scenario</th>
<th>Objective</th>
</tr>
</thead>
<tbody>
<tr>
<td>Computer-aided detection: use of AI to assist in identifying the presence of colorectal polyps and improving adenoma detection rate.</td>
<td>To evaluate the acceptability of AI to assist in the interpretation of medical imagery in detecting colorectal lesions under different bowel preparations and colonic configurations</td>
</tr>
<tr>
<td>Computer-aided characterization: use of AI to classify whether a colonic polyp was adenomatous or hyperplastic.</td>
<td>To evaluate the acceptability of AI to differentiate (without histology) between adenoma (with variable degree of malignant potential) vs hyperplastic polyps (no malignant potent)</td>
</tr>
<tr>
<td>Computer-aided intervention: use of AI in an endoscopy to guide colonoscopic polypectomy.</td>
<td>To evaluate the acceptability of AI to decide which tool to use in assessing the completeness of polypectomy and risk of bleeding, perforation, or both.</td>
</tr>
</tbody>
</table>

AI: artificial intelligence.
Table 2. Survey items used to measure risk perception, acceptance, and trust.

<table>
<thead>
<tr>
<th>Measure</th>
<th>Question text</th>
</tr>
</thead>
<tbody>
<tr>
<td>Risk perception</td>
<td>I expect major risks involved with the artificial intelligence diagnosis.</td>
</tr>
<tr>
<td>Acceptance</td>
<td>Do you believe that machine learning algorithm can, in some cases (as in the one described above), better perform (the task, computer-aided detection, computer-aided characterization, Computer-aided intervention) than human beings?</td>
</tr>
<tr>
<td>Trust</td>
<td>I am ready to try the method myself</td>
</tr>
</tbody>
</table>

Figure 1. Conceptual model of perceived risk, acceptance, and trust on artificial intelligence decision aids.

Statistical Analysis
Statistically significant application pairs were identified by the Mann-Whitney U test (U test) or when there is dependence, the Wilcoxon signed-rank test (Wilcox test). Statistical significance is established at .05. Analyses were conducted in Python using the scipy.stats module (version 1.10.0; the SciPy community), statsmodels module (version 0.13.5), and the Pingouin statistical package (version 0.5.3) or SPSS (version 28; IBM Corp).

Correction for multiple testing was performed using Bonferroni correction, where the statistical threshold ($\alpha$) was divided by the number of tests $n$, such that the adjusted $P$ value threshold is given by $\alpha/n$.

Power Analysis
Our hypothesis is that risk perception, acceptance, and trust will change according to the scenario (detection [CADe], characterization [CADx], or intervention [CADi]), with different levels of invasiveness. Based on an estimated effect size of 0.3 for trust, power, and risk perception with 0.95 power, we can calculate the minimum set of respondents needed to determine any significant differences of a given “size” in response to trust, risk perception, and acceptance measures across scenarios. Since every individual answers scenarios 1 to 3, the differences in the response of every individual can be estimated using a Wilcox test if we compare between pairs of scenarios. The required sample size to pick up a small-moderate effect size (based on Cohen $d$) of 0.3 with a power of 95% is 154. In this study, we have recruited 165 participants, and this should be enough to achieve sufficient statistical power.

Ethical Considerations
This study was approved by the Nanyang Technological University institutional review board (IRB-2022-756). Informed consent was obtained with ability to opt out. Data was anonymized, and no compensation was provided.

Results
Response and Nonresponse Bias
Tracking response rates can help determine the representativeness of a study, but due to the constraints of our institutional review board, we were not allowed to track individual respondents. During the initial phase of the study, we sent the survey to a distribution list of 151 participants with known dates. Applying an approximate 1-month window (October 21, 2022, to November 13, 2022), we obtained 128 responses. Thus, our estimated response rate is ~85% (n=128). While we were analyzing or cleaning up the data, we hoped to get more participants. In the subsequent weeks, we obtained 37 new responses. To compare early and late respondents, we aggregated the first 130 responses (collected between October 21, 2022, and December 29, 2022) as a single group to represent the early respondents and the remaining 35 (collected between January 10 to January 19, 2023) as the late responses. Comparing 130 early respondents against 35 late respondents using a Mann-Whitney U test with a Bonferroni-adjusted $\alpha=.0056$, we found no significant differences for risk, trust, and acceptance across each of the 3 scenarios. This suggests no significant difference between the early and late responses. The lowest obtained $P$ value was .022 (trust in CADx), and the remaining $P$ values were at least .30. Together, we take these
results as a proxy that nonresponder bias is not a strong concern. We also note the overall response rates are rather high; the survey was sent out to various gastroenterology associations as an open invitation, without individual follow-up. It is possible that AI is increasingly seen as transformative and important in the gastroenterology field, but there is not much work on understanding how perspectives on AI lead toward trust and adoption. Hence, invitees feel strongly about the matter and are more inclined to participate in this survey.

**Unidimensionality and Reliability**

Most items in our questionnaire were already used in other questionnaires and can be considered as validated. For the scenario-based questions used in this study, these are novel, as we needed to develop new instruments to explore new topics. Participants had to answer on three 7-point items (not at all to wholeheartedly) whether they accept, trust, and perceive risk on the method presented in each of the scenarios. Unidimensionality and reliability were verified and assured using confirmatory factor analysis and Omega Hierarchical, respectively (see Multimedia Appendix 1 for details).

**Cohort Characteristics**

In total, 165 clinicians participated in the study. The survey completion rate was ~99.40% (n=165). Participants averaged 44.49 (SD 9.65) years, were mostly male (n=116, 70%), and predominantly specialized in gastroenterology (n=153, 92.72%; see Table 3).

The sample comprised gastroenterologists and GI surgeons with varied clinical experience: 93 (56.36%) participants have over 10 years’ experience in practicing gastroenterology and 111 (66.81%) participants were consultants or senior consultants, mostly working in public hospitals (n=110, 66.67%). Most participants reported basic familiarity with AI (n=160, 96.97%; Q1: How familiar are you with AI?). Many were exposed at work, either directly (n=111, 67.27%; Q2: Have you ever used AI in your occupation?) or indirectly (n=112, 67.88%; Q6: Do you personally know other clinicians who use AI at work?).

Participants rated a mean score of 6.00 (SD 0.95) for intending to use AI when it becomes available in their workplace and a score of 5.50 (SD 1.24) for intending to use it to provide services to their patients. Participants rated a mean score of 5.83 (SD 1.37) for intention to use AI routinely in patient care. These figures suggest generally favorable attitudes toward adopting AI.
Table 3. Participant demographics and general characteristics.

<table>
<thead>
<tr>
<th>Participant</th>
<th>Values (N=165), n (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Age (years), mean (SD)</strong></td>
<td>44.49 (9.65)</td>
</tr>
<tr>
<td><strong>Gender</strong></td>
<td></td>
</tr>
<tr>
<td>Male</td>
<td>116 (75.32)</td>
</tr>
<tr>
<td>Female</td>
<td>38 (24.68)</td>
</tr>
<tr>
<td><strong>Country or area</strong></td>
<td></td>
</tr>
<tr>
<td>Australia</td>
<td>3 (1.83)</td>
</tr>
<tr>
<td>Brunei Darussalam</td>
<td>7 (4.27)</td>
</tr>
<tr>
<td>Hong Kong</td>
<td>18 (10.98)</td>
</tr>
<tr>
<td>India</td>
<td>6 (3.66)</td>
</tr>
<tr>
<td>Indonesia</td>
<td>6 (3.66)</td>
</tr>
<tr>
<td>Japan</td>
<td>9 (5.49)</td>
</tr>
<tr>
<td>New Zealand</td>
<td>1 (0.61)</td>
</tr>
<tr>
<td>People’s Republic of China</td>
<td>50 (30.49)</td>
</tr>
<tr>
<td>Philippines</td>
<td>1 (0.61)</td>
</tr>
<tr>
<td>Republic of Korea</td>
<td>2 (1.22)</td>
</tr>
<tr>
<td>Singapore</td>
<td>24 (14.63)</td>
</tr>
<tr>
<td>Taiwan</td>
<td>33 (20.12)</td>
</tr>
<tr>
<td><strong>Main work setting</strong></td>
<td></td>
</tr>
<tr>
<td>Public hospital</td>
<td>110 (67.9)</td>
</tr>
<tr>
<td>Private hospital</td>
<td>28 (17.28)</td>
</tr>
<tr>
<td>Institute of higher learning</td>
<td>18 (11.11)</td>
</tr>
<tr>
<td>Community health center</td>
<td>1 (0.62)</td>
</tr>
<tr>
<td>Other</td>
<td>5 (3.09)</td>
</tr>
<tr>
<td><strong>Current role at work</strong></td>
<td></td>
</tr>
<tr>
<td>Resident</td>
<td>19 (11.8)</td>
</tr>
<tr>
<td>Fellow</td>
<td>19 (11.8)</td>
</tr>
<tr>
<td>Consultant</td>
<td>57 (35.4)</td>
</tr>
<tr>
<td>Senior consultant</td>
<td>54 (33.54)</td>
</tr>
<tr>
<td>Other</td>
<td>12 (7.45)</td>
</tr>
<tr>
<td><strong>Specialty</strong></td>
<td></td>
</tr>
<tr>
<td>Gastroenterology</td>
<td>153 (94.44)</td>
</tr>
<tr>
<td>Colorectal surgery</td>
<td>4 (2.47)</td>
</tr>
<tr>
<td>General surgery</td>
<td>2 (1.23)</td>
</tr>
<tr>
<td>Other</td>
<td>3 (1.85)</td>
</tr>
<tr>
<td><strong>Practicing in specialty (years)</strong></td>
<td></td>
</tr>
<tr>
<td>Less than 5</td>
<td>39 (24.07)</td>
</tr>
<tr>
<td>5-10</td>
<td>30 (18.52)</td>
</tr>
<tr>
<td>11-20</td>
<td>48 (29.63)</td>
</tr>
<tr>
<td>Over 20</td>
<td>45 (27.78)</td>
</tr>
</tbody>
</table>

*a*11 participants did not report their ages or gender.

*b*1 participant did not report their country or area.
Scenario-Based Differentiation

When participants were exposed to three scenarios in medical practice that extend from (1) diagnosing and detecting colorectal polyps (CADe), (2) assessing the nature of pathology of polyps and predict risk of malignancy (CADx), and (3) adopting endoscopic or surgical intervention or removal of the polyps (CADi), clinicians expressed similar risk perceptions across all applications (Figure 2A: Median\textsubscript{CADe}=Median\textsubscript{CADx}=Median\textsubscript{CADi}=4.0; Wilcox\textsubscript{CADe-CADi}: P=.09; Wilcox\textsubscript{CADx-CADi}: P=.44; Wilcox\textsubscript{CADe-CADx}: P=.66).

Figure 2. Gastroenterologists’ attitude toward using AI in the management of colorectal polyps: perceived risk, acceptance, and trust in 3 case scenarios of using AI-assisted colonoscopy in CADe, CADx, and adopting CADi with either surgery or endoscopy. Pairwise tests based on the Wilcoxon test were performed across scenarios. (A) Risk perception across CADe, CADx, and CADi applications. The raincloud plot comprises a 3-panel visualization with a density plot on top revealing density patterns, a box plot in the middle summarizing the median and IQR, and a univariate strip plot on the bottom showing the actual data distribution. No significant pairs were identified. (B) Acceptance across CADe, CADx, and CADi applications. Pairs with statistically significant differences are highlighted by a red connector and an asterisk. (C) Trust across CADe, CADx, and CADi applications. Pairs with statistically significant differences with a P value ≤.02 are highlighted by a red connector and an asterisk. AI: artificial intelligence; CADe: computer-aided detection; CADi: computer-aided intervention; CADx: computer-aided characterization.

Subgroup Analysis for Identification of Confounding Effects and Other Intrinsic Factors

We performed a subgroup analysis to investigate if factors such as gender, years of experience, and practice environment will affect risk perception, acceptance, and trust in AI for gastroenterology practice (Figure 3).

Male and female practitioners held similar risk perceptions. There was good concordance in their risk perception, acceptance, and trust toward using AI in gastroenterology practice (Figure 3A1, 3B1, and 3C1). Male participants tended to be less accepting and trusting, especially in CADi, although this difference is not statistically significant.

Next, we compared practitioners with 10 or less years of clinical experience (n=69) versus experienced practitioners with more than 10 years of clinical experience (n=93). While the overall trends of high acceptance and trust showed no difference between the 2 groups, experienced clinicians exhibited consistently lower risk perception than less experienced ones (Figure 3A2). This observation was statistically significant for all 3 scenarios (CADe: P=9.7\times10^{-6}; CADx: P=1.7\times10^{-6}; CADi: P=3.3\times10^{-4}). We also compared practitioners of the rank senior consultant and consultant (n=111) against residents and fellows (n=38; Figure 3A3, 3B3, and 3C3). The acceptance and trust remained high, and the trend showed a good concordance between the 2 groups. A lower risk perception was found among senior consultants and consultants compared to residents and fellows (CADe: P=.12, CADx: P=.10, and CADi: P=.27). However, the difference is statistically insignificant. The years of experience in clinical practice appeared to have a stronger impact on risk perception than the rank held.

Finally, we compared practitioners from public hospitals with those from private hospitals (Figure 3A4, 3B4, and 3C4). There was no statistically significant difference between private hospital practitioners against their public counterparts, although there was a noticeable difference in CADx on acceptance (Figure 3B4). There was also a lower rate of acceptance and trust in using AI for intervention (CADi) compared to CADe and CADx. Despite not reaching statistical significance, we observed that the spread among private hospital respondents tended to exhibit greater variations. In some instances, the spread appeared to be
bimodal for CADi, suggesting that the private respondents could be a combination of 2 distinct subgroups.

The correlation among risk perception, acceptance, and trust was further analyzed by incorporating the years of experience of the participants by their years of practice in gastroenterology. In all 3 scenarios, there is a moderate correlation between acceptance and trust of AI in detecting polyps (CADe) and characterizing polyps (CADx). The influence of risk perception on acceptance and trust appears to be more diffused: noticeably, when trust and acceptance are both high, and it does not always coincide with low-risk perception.

We first used contingency tables combined with the Fisher exact test to evaluate the impact on the original relationships between trust and acceptance and after introducing risk perception (risk) as an interaction term. This was repeated for each scenario (CADx, CADi, and CADe; Multimedia Appendix 1). Using this approach, we find that after introducing risk perception, the distribution of values still largely follows that of the original data, suggesting that risk does not interact strongly with trust and acceptance. However, this does not mean that risk does not influence these 2 factors. To further investigate, we performed a 2-way ANOVA to further study the influence of risk perception on acceptance and trust. The 2-way ANOVA revealed a statistically significant interaction in CADe ($F_{25}=3.37; P=1.6\times10^{-5}$) but not in CADx ($F_{25}=1.40; P=.12$) and CADi ($F_{36}=1.35; P=.16$). Finally, we performed two sets of regression analyses with (1) acceptance and risk perception as independent variables and (2) acceptance, risk perception, and an interaction term that is the product of acceptance and risk perception (Multimedia Appendix 1). Acceptance had a statistically significant positive influence on trust for all 3 scenarios. Risk perception only has a statistically significant negative impact on trust for the first 2 scenarios (CADe and CADx). When we considered an interaction term, only CADe had a statistically significant impact on trust on all 3 terms. For CADx and CADi, this effect disappeared and only acceptance retained a statistically significant influence on trust. Thus, we believe risk perception has a weak association with trust and acceptance. Taken together, the relationship between trust, acceptance, and risk perception appears complex and is not straightforward.
Figure 3. Subgroup analysis of risk perception, acceptance, and trust stratified by year of experience, seniority (consultant+senior consultant vs fellow+resident), gender, and practicing environment (public vs private hospital). The visualization is a grouped violin plot with split violins. The left and right halves of the violin depict the distributions of 2 samples. If the 2 samples are similar, they will exhibit symmetry on both sides. The median lines for each sample have dashed lines, and these median lines are in turn, bordered by their respective 25th and 75th percentile lines depicted as dotted horizontal lines. Comparisons with statistically significant differences with $P$ value $ \leq 0.0014$ are flagged with a red asterisk. CADe: computer-aided detection; CADi: computer-aided intervention; CADx: computer-aided characterization.
Discussion

Principal Findings

The findings from our study demonstrate that gastroenterologists are generally familiar with AI and were frequently exposed to AI tools in medical settings. This may be because of the introduction of AI-assisted colonoscopy by various industries. In recent years, there are also numerous publications and seminars in the field of gastroenterology mentioning the success of using AI tools in diagnosis, risk prediction, and the treatment of GI conditions [23]. This suggests that they have a keen awareness of AI’s future potential in clinical applications. However, our findings showed that acceptance is not an all-or-nothing choice, but the application or intention to use AI tools varied between different clinical scenarios as well as the nature and impact of AI participation.

When looking at scenario-specific acceptance and trust in AI, the responses vary. Our survey on AI use in detection (CADe), characterization (CADx), and intervention (CADi) of colonic polyps revealed wide acceptance disparity among practitioners (Figure 2). While CADe was more widely accepted, CADi was met with much greater resistance. The 3 AI scenarios that were presented to clinicians in this study varied in the degree of involvement a clinician has in certain procedures. Participants preferred CADi the least. These results agree with our hypothesis that trust, acceptance, and risk perception will change according to the scenario (detection [CADe], characterization [CADx], or intervention [CADi]), with different levels of invasiveness.

In this study, acceptance appeared to have little correlation with the perceived risk level of the procedures. Although certain case scenarios were considered by some as high risk, they do not necessarily warrant low acceptance or trust in using AI. Hence, the findings highlight the intricate relationship between the complexity of AI technologies and their acceptance. One intriguing finding is that participants with more (years of) experience appear to accept the risk and would trust the use of AI more than those who are less experienced. This probably indicates that they see the use of AI as an option or recommendation, instead as an obligation or necessity. Therefore, having more clinical experience may give clinicians greater confidence in their medical expertise and practice, thereby generating more confidence in risk mitigation when new technologies are introduced. Indeed, a study by Lawton et al [24] revealed more experienced doctors were much more at ease with uncertainty.

On the other hand, a general lack of AI familiarization and training in medical education may be one of the reasons that less experienced doctors perceive AI as more risky than regular or traditional practice. Chen et al [25] found that while most physicians and medical students were receptive to the use of AI, most also had concerns about the potential for unpredictable or incorrect results. The same study also stated that respondents were aware of AI’s potential but lacked practical experience and related knowledge. Thus, introducing AI literacy and familiarization training early in medical careers may help mitigate risk aversion and promote responsible AI use in clinical practice. Young doctors are also aware of their education gaps. In a study by Civaner et al [26], medical student respondents acknowledged a gap in “knowledge and skills related to AI applications” (96.2%), “applications for reducing medical errors” (95.8%), and “training to prevent and solve ethical problems that might arise as a result of using AI applications” (93.8%).

Our results suggest that although there is a moderate correlation between trust and acceptance, risk perception appeared invariant suggesting the relationship between trust and acceptance with risk perception is not straightforward and may implicate other factors and interactions than the relationships shown in Figure 1. Indeed, the invariance of risk perception across scenarios against acceptance suggests that there are other factors that influence the acceptance of AI (Figure 2). Among the tested factors, we find that risk acceptance is confounded with years of experience (Figure 3). Future studies should be conducted to better understand other drivers and barriers that influence acceptance, such as the perceived usefulness of using AI and whether AI tools may replace the jobs of clinicians in future practices. Qualitative studies, such as the use of focus group discussions, would also be useful to better understand clinicians’ specific concerns in using AI and the impact of their concerns on the use of AI. Quantitatively, more complex data analysis methods may also be used in the future to understand the causal relationship between various factors and the acceptance of AI. As we proceed into deeper and larger cohort studies investigating trust and acceptance of AI, the development of powerful network methodologies can yield more insight. Indeed, simple statistical learning and even deep learning methods may soon become limited in their ability to explain complex and directed relationships among factors. We believe that causal analysis methods, such as Bayesian Belief Networks will soon become necessary and indispensable for explaining and modeling trust, acceptance, and risk perceptions on medical AI [27].

Limitations

There are limitations in this study. While this study provides invaluable insight into the Asia-Pacific region, we have only captured clinicians’ perspectives despite there being other stakeholders whose voices and opinions matter. This includes nurses, endoscopy assistants, and patients. Future studies should aim to capture their perspectives and understand better how their opinions align or conflict with each other. This will help us navigate complex trust and acceptance issues more realistically and create valuable propositions and effective policies by adopting a multistakeholder perspective into consideration [28]. Participants in this study come from 5 countries with only 165 respondents. The generalizability of the findings can be strengthened by including more clinicians from different backgrounds and regions of practice. In future implementation studies, it may also be worthwhile to examine additional case scenarios such as the management of complicated inflammatory bowel diseases; choice of therapy for GI cancers and GI bleeding; and their corresponding trust, acceptance, and risk perceptions. This additional information will help us better contextualize how risk acceptance, acceptance, and trust change depending on practice.
Conclusions
This study is one of the first to examine risk perception, acceptance, and trust across different scenarios. It is one of the earliest reports of AI risk perception, acceptance, and trust among gastroenterologists, with a unique focus on the Asia-Pacific region. We found that gastroenterologists have, in general, a high acceptance and trust level of using AI-assisted colonoscopy in the management of colorectal polyps. However, this level of trust depends on the application scenario. Moreover, the relationship among risk perception, acceptance, and trust in using AI in gastroenterology practice is not a straightforward correlation. Future studies are required to identify factors that influence the acceptance and trust of using AI in clinical practices.

Acknowledgments
This research or project is supported by the National Research Foundation, Singapore under its AI Singapore Programme (AISG3-GV-2021-009).

Conflicts of Interest
None declared.

Multimedia Appendix 1
Survey questions and supplementary results.

References


The Impact of Expectation Management and Model Transparency on Radiologists’ Trust and Utilization of AI Recommendations for Lung Nodule Assessment on Computed Tomography: Simulated Use Study

Lotte J S Ewals¹, MSc; Lynn J J Heesterbeek², MSc; Bin Yu³, PhD; Kasper van der Wulp¹, MD; Dimitrios Mavroeidis⁴, PhD; Mathias Funk⁵, PhD; Chris C P Snijders⁶, Prof Dr; Igor Jacobs⁷, PhD; Joost Nederend¹, MD, PhD; Jon R Pluyter², PhD; e/MTIC Oncology group⁸

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

Corresponding Author:
Lotte J S Ewals, MSc
Catharina Cancer Institute, Catharina Hospital Eindhoven
Michelangelolaan 2
Eindhoven, 5623 EJ
Netherlands
Phone: 31 40 239 9111
Email: lotte.ewals@catharinizekennhuis.nl

Abstract

Background: Many promising artificial intelligence (AI) and computer-aided detection and diagnosis systems have been developed, but few have been successfully integrated into clinical practice. This is partially owing to a lack of user-centered design of AI-based computer-aided detection or diagnosis (AI-CAD) systems.

Objective: We aimed to assess the impact of different onboarding tutorials and levels of AI model explainability on radiologists’ trust in AI and the use of AI recommendations in lung nodule assessment on computed tomography (CT) scans.

Methods: In total, 20 radiologists from 7 Dutch medical centers performed lung nodule assessment on CT scans under different conditions in a simulated use study as part of a 2x2 repeated-measures quasi-experimental design. Two types of AI onboarding tutorials (reflective vs informative) and 2 levels of AI output (black box vs explainable) were designed. The radiologists first received an onboarding tutorial that was either informative or reflective. Subsequently, each radiologist assessed 7 CT scans, first without AI recommendations. AI recommendations were shown to the radiologist, and they could adjust their initial assessment. Half of the participants received the recommendations via black box AI output and half received explainable AI output. Mental model and psychological trust were measured before onboarding, after onboarding, and after assessing the 7 CT scans. We recorded whether radiologists changed their assessment on found nodules, malignancy prediction, and follow-up advice for each CT assessment. In addition, we analyzed whether radiologists’ trust in their assessments had changed based on the AI recommendations.

Results: Both variations of onboarding tutorials resulted in a significantly improved mental model of the AI-CAD system (informative P=.01 and reflective P=.01). After using AI-CAD, psychological trust significantly decreased for the group with explainable AI output (P=.02). On the basis of the AI recommendations, radiologists changed the number of reported nodules in 27 of 140 assessments, malignancy prediction in 32 of 140 assessments, and follow-up advice in 12 of 140 assessments. The changes were mostly an increased number of reported nodules, a higher estimated probability of malignancy, and earlier follow-up.
The radiologists’ confidence in their found nodules changed in 82 of 140 assessments, in their estimated probability of malignancy in 50 of 140 assessments, and in their follow-up advice in 28 of 140 assessments. These changes were predominantly increases in confidence. The number of changed assessments and radiologists’ confidence did not significantly differ between the groups that received different onboarding tutorials and AI outputs.

**Conclusions:** Onboarding tutorials help radiologists gain a better understanding of AI-CAD and facilitate the formation of a correct mental model. If AI explanations do not consistently substantiate the probability of malignancy across patient cases, radiologists’ trust in the AI-CAD system can be impaired. Radiologists’ confidence in their assessments was improved by using the AI recommendations.

**KEYWORDS**
application; artificial intelligence; AI; computer-aided detection or diagnosis; CAD; design; human centered; human computer interaction; HCI; interaction; mental model; radiologists; trust

**Introduction**

**Background**
Lung cancer is one of the leading causes of cancer-related deaths worldwide [1]. Early detection of lung cancer is essential to provide curative treatment and improve survival. However, detecting and diagnosing lung cancer using computed tomography (CT) scans can be challenging. On CT scans, early lung cancer can be seen as a small nodule. However, these nodules can also be benign. The risk of malignancy depends on various patient factors and lung nodule features, such as the morphology, size, and number of lung nodules. Nodules that are challenging to detect can, for instance, be small, and their perceptibility might be hampered by their location close to normal lung tissue that is visually similar on a CT scan, such as blood vessels or bronchi [2-5]. As a result, radiologists may overlook or misdiagnose lung nodules on CT scans. A previous study showed that radiologists missed 15% of all lung cancer cases on screening CT scans. Of these missed cancers diagnoses, 35% were not visible on the scan, 50% were not detected by the radiologist, and 15% were detected but not diagnosed as cancer [6].

A recent approach to improve the detection and diagnosis of lung nodules on CT scans is the use of artificial intelligence (AI) models. Diagnostic assistance from AI models that provide recommendations for radiologists is referred to as AI-based computer-aided detection or diagnosis (AI-CAD) [7]. Many studies have been published on AI models for assessing lung nodules on CT scans, showing promising performance with sensitivities for detection of up to 98.1% and a mean of only 2 false-positives (FPs) per scan [8,9].

Although many AI models and AI-CAD systems have been developed, few are used in clinical practice. Although most studies on AI for lung nodule assessment focus on the development and stand-alone performance of AI models [8,10,11], few studies have focused on user interaction with AI models in the clinical context beyond the theoretical level [12-16]. However, human-AI interaction is essential to enable radiologists to comprehend and effectively use AI recommendations in their tasks, ultimately achieving the highest levels of diagnostic quality and efficiency.

Trust is of great importance in the interactions and collaborations between radiologists and AI-CAD systems [15,17-20]. Trust influences the end users’ level of reliance on AI recommendations, and hence, it influences the performance of AI-assisted end users [18,19]. If the user has very little trust in the system, the potential benefits of AI-CAD will be reduced because of disuse, whereas too much trust in the system leads to overreliance and can result in mistakes that would not have been made without using the AI-CAD system [15,18].

Trust is a dynamic process. Trust changes over time and across situations and is influenced by many factors. For example, trust varies based on the reliability of the AI system, the design of the system, the personal characteristics of the user, prior interactions and experience, and moderating factors such as workload and sociocultural context [18,21-25]. Some of these factors can be influenced through the design of the system, with the aim of achieving the formation of appropriate trust. Trust calibration refers to interventions that facilitate the formation of an appropriate trust level by aligning a person’s trust in the AI with the capabilities of the AI [26,27]. In this study, we introduced 2 instruments aimed at appropriate trust calibration at different time points of use. First, an onboarding tutorial aimed to set the right expectations before initial use. Second, AI model explainability as an information cue available to clinical users during use to judge the credibility of the arguments underpinning the AI model prediction.

We aimed to assess whether radiologists’ trust in AI-CAD systems and their use of AI recommendations in lung nodule assessments on CT scans were affected by different onboarding tutorials and by different levels of AI model explainability.

**Theoretical Argumentation**

**Trust Definitions**
Different definitions and measures exist for trust [15]. In this study, we considered trust from 2 complementary perspectives, a cognitive perspective and a behavioral perspective [23].

From the cognitive perspective, we explored the users’ mental model and psychological trust. The mental model represents a person’s “static knowledge about the system: its significant features, how it functions, how different components affect others, and how its components will behave when confronted with various factors and influences” [24]. In short, the mental
model is the user’s understanding of the AI system. A correct mental model is expected to contribute to appropriate trust calibration between the user’s trust in an AI system and the trustworthiness of the system [25]. User’s psychological trust refers to “the extent to which a user is confident in, and willing to act on the basis of, the recommendations, actions, and decisions of an artificially intelligent decision aid” [28]. Because radiologists gain experience and learn through the process of assessing CT cases with the AI-CAD tool and actually see what the system is capable of, they are expected to have an improved mental model of (hypothesis 1a) and psychological trust in (hypothesis 1b) the AI-CAD system after using the AI-CAD system compared with before using the system.

However, holding a positive attitude toward the AI-CAD system does not mean that the user will also act in line with its recommendations. Therefore, we also adopted a behavioral lens by examining whether trust was reflected in the use of the AI recommendations (reliance and compliance) and the corresponding impact on decision outcomes [29,30]. The decision of whether radiologists use AI recommendations depends not only on their overall trust in the AI-CAD system but also on their agreement with the specific AI recommendations for a given case. As the AI recommendations function as a second reader, it is expected that radiologists’ confidence in their assessments will be higher when they are assisted by AI-CAD than without assistance (hypothesis 2).

Onboarding Tutorials

Research on how to ensure that radiologists have appropriate expectations of the system’s capabilities and limitations is limited [27]. As suggested by Cai et al [31], when clinical practitioners are first introduced to an AI system, a human-AI onboarding process can be crucial for them to determine how they will partner with AI in practice. Therefore, an onboarding tutorial to inform radiologists about the capabilities and limitations of the AI-CAD system is expected to improve radiologists’ mental model of (hypothesis 3a) and psychological trust in the AI-CAD system (hypothesis 3b).

Moreover, critical reflection on one’s experience is essential for developing competence and self-awareness [32]. Hence, it is hypothesized that critical reflection and feedback built through a reflective onboarding tutorial will lead to a more improved mental model of (hypothesis 4a) and psychological trust in (hypothesis 4b) the AI-CAD system than an informative onboarding tutorial. Furthermore, it is expected to be easier for radiologists to understand whether an AI suggestion should be followed because of their understanding of the AI-CAD system from reflective onboarding, especially when they are not fully sure of their own assessment. Therefore, it is expected that radiologists who receive reflective onboarding will use the AI recommendations more often than radiologists who receive informative onboarding (hypothesis 5).

Levels of AI Model Explainability

In addition, radiologists are expected to better judge whether they can trust an AI recommendation when the AI model discloses the reasoning behind its recommendations (explainable AI models) compared with black box models. Hence, it is hypothesized that after using the AI-CAD system, radiologists assisted with explainable AI output have an improved mental model of (hypothesis 6a) and psychological trust in (hypothesis 6b) the AI-CAD system than radiologists assisted with black box AI output. Because radiologists can see the reasoning behind the recommendations when receiving explainable AI output, it is expected that they will use the AI recommendations more often than radiologists assisted with black box AI output (hypothesis 7).

Methods

Overview

We tested the hypotheses using a 2×2 repeated-measures quasi-experimental design: informative versus reflective onboarding tutorial and black box versus explainable AI output. In this simulated use study, we aimed to realistically mimic clinical practice [33,34]. Realistic clinical simulations allow participants to engage with the setup in real-world clinical scenarios and encourage participants to authentically execute the study as if they are performing their clinical work.

Prototype

Image Viewer

A medical image–viewing prototype was developed to enable radiologists to assess incidental lung nodules on cardiac CT scans with and without the assistance of an AI-CAD system. The AI recommendations were implemented as a second reader, allowing the radiologist to first assess the cases independently. The interface was designed based on the literature, brainstorming, and feedback sessions with radiologists and design specialists and was iteratively optimized for the 2 variations of onboarding tutorials (reflective vs informative) and 2 variations of AI outputs (black box vs explainable). The final user interface is shown in Figure 1. We aimed to realistically simulate the radiologists’ clinical setup to facilitate proper engagement of the participants with the task of lung nodule assessment. The user setup was designed to simulate clinical practice as realistically as possible. The developed interface was shown to the radiologists on a monitor, which was placed in a separate silent room. This room was inside the hospital, and lights could be dimmed if the radiologists preferred it, comparable with their own working space. Similar to the picture archiving and communication system used in clinical practice to assess CT scans, radiologists could scroll through the images, zoom in, measure, and change the windowing level between the soft tissue and lung setting using a computer mouse.

https://ai.jmir.org/2024/1/e52211 JMIR AI 2024 | vol. 3 | e52211 | p.175
**Clinical Data**

To further increase study engagement and realism, the use scenarios were based on real-world patient cases. We retrospectively selected 10 CT angiography scans with incidental pulmonary nodules from a large Dutch clinical hospital. Scans acquired between 2008 and 2015 were used because the 5-year outcomes of these patients are known: whether they developed lung cancer. An expert radiologist selected the cases for this study. Of the 10 selected scans, we used 3 for onboarding and 7 for testing the impact of the design interventions. All CT scans were performed on patients with lung cancer. By selecting the 7 CT cases, we aimed to obtain a diverse mix of assessment complexity by including both lower and higher suspicious nodules (based on size, spiculation, and solidity) and nodules at easier and more difficult locations (such as against the veins or pleura). The characteristics of the 7 CT cases and the findings of the AI model for these cases are presented in Multimedia Appendix 1.

**AI Model**

To detect and estimate the malignancy of lung nodules on the CT scans, the pretrained AI framework developed by Trajanovski et al [35] was applied. This framework relies on a 2-stage process, where the first stage performs nodule detection and the second stage assigns a malignancy probability to the detected nodules. Among the validated nodule detectors, the best performance was achieved by the nodule detector developed by Liao et al [36]. This nodule detector is based on deep learning models, more precisely, convolutional neural networks. The nodules detected by the nodule detector are provided as input to the second stage of the framework that assigns the cancer malignancy probabilities. The second stage of the framework is based on a convolutional neural network that was trained using the publicly available National Lung Screening Trial data set [37].

During inference, the model takes a CT scan as input and automatically produces a list of nodule locations (x,y,z), their radii, and malignancy probabilities. The prototype, described previously, ensures that this information is displayed intuitively to the clinicians. The article by Liao et al [36] provides all the relevant details regarding the training process and performance validation.

In this study, the AI model proposed by Trajanovski et al [35] was used without any additional fine-tuning. Specifically, the model weights remained unchanged. The sole adjustment involved calibrating (or rescaling) the output of the model to accommodate the changed distribution of malignant cases (Multimedia Appendix 2 [35,38,39]).

**AI Recommendations**

The AI model recommendations were provided using 4 information cues (Multimedia Appendix 3):

1. Detected nodules (shown by target mark directly on the CT scan)
2. Benign or malignant classification per nodule (malignant nodules are highlighted in orange color)
3. Model confidence in the benign or malignant classification (shown as the negative predictive value [NPV] or positive predictive value [PPV] score and an intuitive icon representing high, medium, or low confidence)
4. In the explainable AI output variant: nodule features serving as explanations for the classification

AI nodule detection and benign or malignant classification (cues 1 and 2) were obtained using the described AI model [35]. The number of lung nodules detected by the AI model varied...
between 1 and 5 per scan. The AI model found at least one true-positive lung nodule in each case and found one or more FP nodules in 4 of 7 cases. For more information about the AI findings, see Table S1 in Multimedia Appendix 1.

Confidence in the malignancy classification (cue 3) was given by means of PPVs for malignant predictions, indicating the probability that nodules with malignant predictions were actually malignant, and by means of NPVs for benign predictions, indicating the probability that nodules with a benign prediction were actually benign. The PPV was 0.25 (low confidence), 0.30 (medium confidence), or 0.38 (high confidence), and the NPV was 0.94 (low confidence), 0.97 (medium confidence), or >0.99 (high confidence; for an explanation of how the PPVs and NPVs were calculated, see Multimedia Appendix 2). In addition, confidence was shown by means of a small bar graph, indicating low, medium, or high model confidence.

Two levels of AI transparency were tested: black box AI output and explainable AI output. Black box output indicates that radiologists did not see what the malignancy estimation was based on. The explainable AI output variant provided the same information as the black box AI output variant and additionally showed the characteristics of the lung nodules (cue 4); this information was expected to help in understanding and interpreting the predictions of the AI-CAD system (Figure 1, right column). For each lung nodule, the following lung nodule characteristics were provided: long axis diameter, solidity, margin characteristics, and location. The nodule characteristics were not provided by the AI model and were therefore realistically simulated, which is in agreement with related research [40] via manual annotation by 2 expert radiologists in consensus. However, the participants were not aware of the simulation; therefore, from the radiologists’ perspective, the characteristics were AI generated as well [41]. For an overview of the information cues for the AI recommendations, see Multimedia Appendix 3.

Onboarding Tutorials

Two variations of onboarding tutorials were designed: informative onboarding and reflective onboarding. During informative onboarding, radiologists passively received a stepwise introduction of the AI capabilities and common pitfalls so that they could acquire a realistic mental model of the system (Figure 2). The AI model’s capabilities and pitfalls were illustrated in the onboarding tutorial with 3 CT scans that showed obvious cancer cases, FP nodules, and false-negative nodules. For an overview of all implemented questions and explanations, see Multimedia Appendix 3. During reflective onboarding, radiologists additionally engaged in active reflection. They received cognitive feedback on 4 questions that they had to answer to check whether their mental model of the AI-CAD system was correct.

Study Protocol

For this study, physicians were eligible for participation if they were radiologists, nuclear radiologists, or radiology residents. We will refer to the participants as radiologists. Several effects were to be tested; we used a power of 80%. For the mental model differences between radiologists, we based our sample size calculation on a comparison of means of 2 versus 3 (SD 0.5). This led to a necessary sample size of 12 radiologists. For the psychological trust differences, we based the sample size calculation on a comparison of means of 0.5 versus 0.75 (SD 0.1). This resulted in a sample size of at least 8 radiologists.
The differences in the use of AI recommendations were based on a comparison of proportions in the order of magnitude of 30% versus 10%. This leads to a necessary sample size of 124 comparisons if we assume that the intraclass coefficient is low. Eventually, 20 radiologists were included in this study, all of whom assessed 7 CT scans for a total of 140 recommendations [42]. In this 2x2 repeated-measures design, the radiologists were divided into 4 groups, each of which consisted of 5 radiologists. After onboarding in one of the 2 conditions, using 3 CT scans, each radiologist assessed the 7 CT scans. In addition to the CT scans, each patient’s age and gender were provided because radiologists also use the patient context when they assess CT scans in clinical practice. First, the radiologists assessed the scans without observing the AI output. They reported the nodules they detected, estimated the malignancy probability for the patient case (not per nodule, unlike the AI model), and provided follow-up advice. Subsequently, the AI recommendations were presented, and the radiologists could adjust their initial assessments. The nodules detected by AI and the AI malignancy estimations might trigger the radiologists to change their initial assessments. This process is visualized in the flow diagram in Figure 3.

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

**Measures for Trust**

To evaluate the effects of the 2 types of AI onboarding tutorials and the 2 levels of explainability of AI outputs on radiologists’ trust in AI and their use of AI recommendations, participants were requested to complete questionnaires on 3 aspects: the radiologists’ mental model of the AI-CAD system’s capabilities and pitfalls, psychological trust in the AI-CAD system, and the use of AI recommendations. These questionnaires were completed at different time points, as schematically shown in Figure 4.

**Figure 4.** Overview of the flow of the experiment with the questionnaires at different time points. AI: artificial intelligence; CT: computed tomography.

**Mental Model**

The mental model questionnaire measured the radiologists’ understanding of the AI capabilities and limitations to uncover whether their expectations of the AI-CAD system were appropriate. Of the 11 questions in this questionnaire, 5 questions were related to nodule detection and 6 were related to malignancy prediction (see the full questionnaire in Multimedia Appendix 4). Questions could be answered with yes, no, or I do not know. Depending on whether the assessment was correct as compared with the true AI capabilities, a score
of 1 (correct) or 0 (incorrect or I do not know) was assigned per question, resulting in summed scores between 0 and 11. A higher score implies a better understanding of the AI capabilities. The mental model was measured before onboarding, after onboarding, and after assessing the 7 CT scans.

**Psychological Trust**

To measure the radiologists’ psychological trust in the AI-CAD system, a questionnaire was derived from the study by Ashoori and Weisz [43] and adapted to fit this study (see the full questionnaire in Multimedia Appendix 4). This questionnaire examined overall trustworthiness, reliability, technical competence, and personal attachment. An example of a statement is “This model is trustworthy.” The 12 statements about the AI model had to be answered with a score between 1 (strongly disagree) and 5 (strongly agree). For the negatively phrased questions, scores were reversed for the data analysis so that for all questions, a higher score reflected more trust in the AI-CAD system. Subsequently, the scores for the 12 questions were averaged. The psychological trust of each participant was measured before onboarding, after onboarding, and after assessing the 7 CT scans.

**Use of AI Recommendations**

To evaluate the radiologists’ use of the AI recommendations, their assessments and confidence in their assessments—first without and then with AI assistance—were recorded in a questionnaire. AI recommendation use was measured at 3 assessment levels: number of detected nodules, malignancy probability, and follow-up advice. Therefore, the questionnaire included questions about the number of found nodules, the malignancy probability (at the patient level) as a percentage, and the follow-up advice according to the Fleischner guidelines [44]. The follow-up advice had to be scored with a score of 1 (consider CT at 3 months, positron emission tomography–CT, or tissue sampling), 2 (CT at 3-6 months), 3 (CT at 6-12 months), 4 (CT at 12 months), or 5 (no routine follow-up). A lower score indicated earlier follow-up. In addition, the confidence of the given answers at each assessment level had to be rated with a score between 1 (not confident at all) and 5 (very confident). The complete questionnaire is provided in Multimedia Appendix 4. Participants were requested to complete this questionnaire while assessing without AI assistance and with AI assistance for each CT case.

**Analyses**

**Mental Model and Psychological Trust**

Changes in the mental model and psychological trust were assessed by comparing the scores before and after onboarding, and the scores after onboarding and at the end of the test, that is, after assessing all 7 CT scans. These changes were assessed for all radiologists together, for the 2 onboarding tutorial groups separately, and for the 2 AI output groups separately. The changes in scores were compared between the 2 onboarding tutorial groups and between the 2 AI output groups to analyze whether the types of onboarding tutorials and level of AI explainability influenced radiologists’ initial trust and maintenance of trust during CT assessment. In addition, we analyzed whether the changes in mental model and psychological trust scores were influenced by any of the following characteristics of the radiologists: age, gender, years of experience, how often they assessed lungs on CT as part of their job, how eager they were to try new information technologies, and how frequently they used AI-CAD tools.

**Use of AI Recommendations**

The use of AI recommendations was assessed by analyzing the number of cases in which radiologists adjusted the number of found nodules, the malignancy probability, and the follow-up advice after viewing the AI-CAD recommendations. In addition, we analyzed whether the radiologist’s confidence in the assessments of the number of nodules, the malignancy prediction, and the follow-up advice changed after viewing the AI recommendations and whether their confidence increased or decreased. The use of AI recommendations and the impact on radiologists’ confidence were compared between the groups of onboarding tutorials and between the groups of AI output.

**Secondary Analyses**

**Additional Analyses and Use of AI Recommendations**

In addition, the impact of agreeing or disagreeing with the AI detected nodules was evaluated. We analyzed whether the use of AI recommendations and radiologists’ confidence in their assessments were affected by 2 factors: first, whether the same or different nodules were found by the AI as compared with the radiologist and, second, whether the radiologist changed the number of reported nodules after seeing the AI recommendations.

**Correctness of Follow-Up Advice**

Furthermore, to evaluate whether AI-CAD assistance resulted in improved clinical assessment, we analyzed whether the radiologists selected the correct follow-up advice more often with or without the AI recommendations. For each case, the correct follow-up according to the Fleischner criteria was retrospectively determined by 2 expert radiologists in consensus and used as reference follow-up advice. The follow-up recommendations provided by the radiologists were compared with the reference follow-up advice, and we analyzed whether AI assistance resulted in more accurate follow-up advice.

**Statistical Analyses**

**Mental Model and Psychological Trust**

Differences between the mental model scores and psychological trust scores of the radiologists at different time points were analyzed using the Wilcoxon signed rank test. Differences between the mental model scores and psychological trust scores of the groups with informative and reflective onboarding tutorials and of the groups with black box and explainable AI output were statistically analyzed using Mann-Whitney U tests. To control for heterogeneity, we tested whether radiologists’ characteristics influenced the mental model scores and psychological trust scores at different time points and over time by performing multiple linear regression analyses.

**Use of AI Recommendations**

Multilevel logistic regression analyses were performed to assess whether the type of onboarding tutorial or level of explainability...
of the AI output influenced the use of the AI recommendations and the radiologists’ confidence in their assessments. To control for potential impact on the outcomes by other factors (exclusively the same nodules found by radiologists and AI model, change in number of reported nodules, age, gender, years of experience, how frequently they assess lungs on CT, how eager they are to try new information technologies, and how frequently they used computer-aided detection tools), these factors were included in the multilevel regression analyses as well. The same analysis scheme was used for all multilevel logistic regression analyses. First, an empty model was run to identify the variance at the individual level. The second regression analysis also considered the variants of onboarding tutorials and AI output. Third, whether the same nodules were found by AI and the radiologist exclusively and whether they made changes in the number of reported nodules were added. The final analysis also included different CT scans and radiologists’ characteristics.

A P value of <.05 was considered statistically significant. All analyses were performed using Stata (version 17; StataCorp).

Ethical Considerations
This study was approved by the Internal Committee for Biomedical Experiments of Philips (number ICBE-S-000204) and conducted in accordance with the Declaration of Helsinki (as revised in 2013). Written informed consent was obtained from the participating clinicians.

Results

Participants
In total, 20 physicians from 7 Dutch hospitals participated in this study. Of the 20 participants, 16 were radiologists (median 10.5, range 1-32 years of experience as a specialist), 1 was a nuclear radiologist (2 years of experience in assessing lung CT scans), and 3 were radiology residents (median 2, range 1-5 years of residency). Of the 16 radiologists, 8 (50%) specialized in thoracic radiology. The male-to-female ratio was 50:50. Of the participants, 25% (5/20) were aged between 26 and 35 years, 35% (7/20) were aged between 36 and 45 years, 20% (4/20) were aged between 46 and 55 years, and 20% (4/20) were aged between 56 and 65 years.

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

After onboarding, the mental model score of the radiologists was significantly higher than that before onboarding (P<.001). The mean scores were 5.7 (SD 2.0) before onboarding and 8.6 (SD 1.9) after onboarding, which supports hypothesis 3a. Both informative (P=.01) and reflective (P=.01) onboarding resulted in significantly higher mental model scores. These improvements did not significantly differ between the groups; therefore, hypothesis 4a is not supported. At the end of the test, the mental model scores did not differ significantly from the scores after onboarding in any of the groups, which does not support hypothesis 1a and hypothesis 6a.

Considering all radiologists together, the psychological trust scores did not change significantly over time; therefore, hypotheses 1b and 3b are not supported. Between the 2 variations of onboarding tutorials, no significant differences in psychological trust scores were observed, and therefore, hypothesis 4b is not supported. In the group that received explainable AI output, psychological trust at the end of the test was significantly lower than that after onboarding (P=.02), which interestingly contradicts hypothesis 6b. In the group that received black box AI output, there was no significant change in psychological trust. Changes in psychological trust scores between after onboarding and at the end of the test were significantly different between the black box output and explainable AI output groups (P=.03). All P values can be found in Multimedia Appendix 5.

None of the tested characteristics of radiologists significantly predicted the mental model scores or the psychological trust scores at the different time points nor did they significantly predict the changes over time.
Figure 5. Boxplot showing the (A) mental model scores and (B) psychological trust scores before and after onboarding and at the end of the test using either informative or reflective onboarding tutorials and either black box or explainable artificial intelligence (AI) output. The cross shows the mean value; the horizontal line inside the box indicates the median value; the lower and higher boundaries of the box indicate the first and third quartiles; the whiskers indicate the minimum and maximum values; and outliers are indicated by colored dots. Only significant differences are mentioned. *Significant difference between time points. **Significant difference in the change over time between the black box and explainable AI output groups.

Use of AI Recommendations

After viewing the AI outcomes, the radiologists adjusted their found nodules in 27 of 140 assessments, their estimated probability of malignancy in 32 of 140 assessments, and their follow-up advice in 12 of 140 assessments (Figure 6). Radiologists predominantly added nodules (23 of 27 changed cases), increased the probability of malignancy (24 of 32 changed cases), and shortened the recommended follow-up period (eg, from CT at 6-12 months to CT at 3-6 months; 8 of 12 changed cases). The empty model, which included no predictor variables, revealed that regarding whether radiologists made changes, approximately 3% of the variance in the outcome variable was attributable to differences between radiologists. For changes in malignancy prediction and follow-up advice, this attributable variance was approximately 20% and 7%, respectively. This indicates that there is some variability in the outcome, which can be explained by the individual radiologists. Radiologists’ assessments were not significantly impacted by the type of onboarding tutorial or by the type of AI output; therefore, hypotheses 5 and 7 are not supported. All outcomes of the multilevel regression analyses can be found in Multimedia Appendix 6.

At all levels of assessment, radiologists’ confidence in the assessments (n=140) predominantly increased after viewing the AI-CAD recommendations (in found nodules [75/82, 91%] of all changed assessments, in malignancy probability [42/50, 84%], in follow-up advice [22/28, 79%]; Figure 7), which supports hypothesis 2. The multilevel regression analysis revealed that in the empty model without predictor variables, approximately 20% of the total variance in the changed confidence in detected nodules was attributed to differences between radiologists. Regarding the changed confidence in malignancy prediction and follow-up advice, this attribution of the total variance was 10% and 7%, respectively. The radiologists’ confidence in their assessments was not significantly affected by the type of onboarding tutorial but was affected by the type of AI output after controlling for whether the AI model found the same or different nodules as the radiologist without AI assistance (first model: $\beta$=0.143; $P$=.16; second model: $\beta$=0.167; $P$=.04; third model: $\beta$=0.207; $P$=.02).
See Multimedia Appendix 6 for all outcomes of the multilevel regression analyses.

Figure 6. Bar graph showing the changes in the radiologist’s computed tomography assessments; (A) Reported nodules, (B) Malignancy probability, (C) Follow-up advice after viewing the recommendations from the artificial intelligence–based computer-aided detection or diagnosis using either informative or reflective onboarding tutorials, and either black box or explainable artificial intelligence (AI) output. No significant differences between the onboarding and AI output groups resulted from the multilevel regression analyses.
Figure 7. Bar graph showing the changes in the radiologist’s confidence in their assessments; (A) Confidence reported nodules, (B) Confidence malignancy probability, (C) Confidence follow-up advice after viewing the recommendations from the artificial intelligence–based computer-aided detection or diagnosis using either informative or reflective onboarding tutorials, and either black box or explainable artificial intelligence (AI) output.

*The multilevel regression analysis showed a significant difference between the 2 groups according to the number of changed radiologists’ confidence (orange+green) in their assessment after using the artificial intelligence–based computer-aided detection or diagnosis system.
Secondary Outcomes

Post Hoc Analyses Regarding the Use of AI Recommendations

In 26 of 140 assessments, the same nodules exclusively had been found by the AI model and the unassisted radiologist. In these cases, radiologists changed the number of nodules less frequently than when different nodules had been found (second model: β=-0.245; P=.003 third model: β=-0.437; P=.001; Multimedia Appendix 6).

In 27 of 140 assessments, radiologists changed the number of nodules when using AI assistance. In the cases in which the radiologists did not change the number of nodules, the radiologists’ confidence in their malignancy prediction changed more often, mostly increased, than in the cases in which the radiologists did change the number of found nodules (second model: β=0.369; P<.001; third model: β=0.283; P=.001; Multimedia Appendix 6). Whether the number of nodules was changed also significantly influenced radiologists’ confidence in their follow-up advice, but this was probably related to some radiologists’ characteristics, as this effect disappeared after controlling for such characteristics (second model: β=0.277; P=.02; third model: β=0.154; P=.23).

Correctness Follow-Up Advice

Without AI assistance, the radiologists provided the correct follow-up advice according to the Fleischner criteria in 94 of 140 assessments (Table 1). Mostly, the correct follow-up advice was provided for CT cases 1, 3, 5, and 7, whereas most of the incorrect follow-up advice concerned CT cases 2, 4, and 6. With AI assistance, radiologists provided correct follow-up advice in 100 of 140 assessments. In 12 cases, the follow-up advice was changed after viewing the AI results. In 7 of these 12 cases, correct follow-up was provided after seeing the AI results. In 1 case, correct follow-up advice that was given initially was changed to incorrect follow-up advice after seeing the AI results. In 3 cases, the changed follow-up advice was still not correct but closer to the correct follow-up advice, and in the remaining case, the changed follow-up advice was further from the correct follow-up advice.

Discussion

Principal Findings

Our study demonstrated that onboarding is of great importance because the radiologists’ mental model of the AI-CAD system was significantly more accurate after onboarding. This finding implies that after onboarding, radiologists had a better understanding of the capabilities and limitations of the AI-CAD system, which is important for using the AI recommendations correctly. In addition, the importance of onboarding was emphasized by the fact that the mental model did not become more accurate through the actual use of the AI-CAD system. A study by Lam Shin Cheung et al [45] supports the need for onboarding.

We hypothesized that reflective onboarding would result in a more appropriate level of trust than informative onboarding, as radiologists in the reflective onboarding group were triggered to actively engage in cognitive reflection and receive feedback on their mental model. However, this hypothesis was not supported because the increases in mental model scores of radiologists in the reflective onboarding group did not significantly differ from those in the informative onboarding group. This unexpected finding might be explained by the high level of clarity of the explanations provided during both informative and reflective onboarding, because of which the reflection had no significant added value. Alternatively, participating radiologists might possess a natural tendency to engage in cognitive reflection even if the system does not actively trigger them to do so.

Another unexpected finding was that explainable AI output resulted in a significant decrease in psychological trust (P=.02) during the use of the AI-CAD system for assessing the 7 CT scans, which was not the case in the group that received black box AI output (Figure 5). Apparently, users can become insecure about the reliability of AI-CAD when they receive explanations. On the basis of feedback from the participating radiologists, we know that some radiologists observed that the AI-CAD system provided different malignancy predictions for similar nodules with the same visual characteristics provided such as size and morphology. These discrepancies raised questions about why

Table 1. Correct follow-up advice provided by the radiologists.

<table>
<thead>
<tr>
<th>CT cases (number of assessments)</th>
<th>All (n=140)</th>
<th>CT1 (n=20)</th>
<th>CT2 (n=20)</th>
<th>CT3 (n=20)</th>
<th>CT4 (n=20)</th>
<th>CT5 (n=20)</th>
<th>CT6 (n=20)</th>
<th>CT7 (n=20)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Correct follow-up advice given without AI assistance, n (%)</td>
<td>94 (67)</td>
<td>20 (100)</td>
<td>6 (30)</td>
<td>17 (85)</td>
<td>6 (30)</td>
<td>20 (100)</td>
<td>7 (35)</td>
<td>18 (90)</td>
</tr>
<tr>
<td>Correct follow-up advice given with AI assistance, n (%)</td>
<td>100 (71)</td>
<td>20 (100)</td>
<td>7 (35)</td>
<td>17 (85)</td>
<td>7 (35)</td>
<td>20 (100)</td>
<td>9 (45)</td>
<td>20 (100)</td>
</tr>
<tr>
<td>Changed follow-up advice after using AI assistance, n (%)</td>
<td>12 (9)</td>
<td>0 (0)</td>
<td>2 (10)</td>
<td>1 (5)</td>
<td>5 (25)</td>
<td>0 (0)</td>
<td>2 (10)</td>
<td>2 (10)</td>
</tr>
<tr>
<td>Wrong→correct</td>
<td>7 (58)</td>
<td>0 (0)</td>
<td>1 (50)</td>
<td>0 (0)</td>
<td>2 (40)</td>
<td>0 (0)</td>
<td>2 (100)</td>
<td>2 (100)</td>
</tr>
<tr>
<td>Wrong→better (still wrong, but closer to correct follow-up)</td>
<td>3 (25)</td>
<td>0 (0)</td>
<td>1 (50)</td>
<td>1 (100)</td>
<td>1 (20)</td>
<td>0 (0)</td>
<td>0 (0)</td>
<td>0 (0)</td>
</tr>
<tr>
<td>Wrong→worse (still wrong, even further from correct follow-up)</td>
<td>1 (8)</td>
<td>0 (0)</td>
<td>0 (0)</td>
<td>0 (0)</td>
<td>1 (20)</td>
<td>0 (0)</td>
<td>0 (0)</td>
<td>0 (0)</td>
</tr>
<tr>
<td>Correct→wrong</td>
<td>1 (8)</td>
<td>0 (0)</td>
<td>0 (0)</td>
<td>0 (0)</td>
<td>1 (20)</td>
<td>0 (0)</td>
<td>0 (0)</td>
<td>0 (0)</td>
</tr>
</tbody>
</table>

CT: computed tomography. AI: artificial intelligence.
nODULES WITH SIMILAR CHARACTERISTICS HAD DIFFERENT MALIGNANCY PROBABILITIES. IN FACT, THIS KEY ASPECT STILL FELT LIKE A BLACK BOX TO THE PARTICIPANTS. APPARENTLY, PROVIDING MORE TRANSPARENCY, WHICH ENABLES RADIOLOGISTS TO OBSERVE INCONSISTENCIES IN THE AI PREDICTIONS, CAN DECREASE THE RADIOLOGISTS’ TRUST IN THE AI-CAD SYSTEM. HOWEVER, THIS DECREASE IN TRUST MIGHT BE APPROPRIATE BECAUSE THE AI MODEL’S PERFORMANCE MIGHT BE SUBOPTIMAL AND INCONSISTENT.


ANOTHER IMPORTANT FINDING IS THAT RADIOLOGISTS BECAME MORE CONFIDENT IN THEIR ASSESSMENTS AFTER USING THE AI RECOMMENDATIONS. THIS CHANGE MIGHT BE EXPLAINED BY THE FACT THAT THE AI-CAD SYSTEM PROVIDES AN EXTRA CHECK, WHICH REDUCES THE LIKELIHOOD OF NODULES BEING OVERLOOKED. HENCE, IT PROVIDES RADIOLOGISTS WITH A SENSE OF SAFETY THAT INCREASES THEIR CONFIDENCE, REGARDLESS OF WHETHER THEY AGREE WITH THE AI OUTPUT.

THE FOLLOW-UP ADVICE WAS ADJUSTED BY THE RADIOLOGISTS AFTER VIEWING THE AI RESULTS IN ONLY 12 OF 140 ASSESSMENTS, WHEREAS THE NUMBER OF OBSERVED NODULES AND THE MALIGNANCY PROBABILITIES WERE CHANGED MORE OFTEN (27/140, 19.3% ASSESSMENTS AND 32/140, 22.9% ASSESSMENTS, RESPECTIVELY). THIS FINDING CAN BE EXPLAINED BY THE FACT THAT FOLLOW-UP ADVICE IS PROMINENTLY AFFECTED BY THE MOST SUSPICIOUS NODULE. CONSEQUENTLY, AN AI-CAD FINDING OF AN ADDITIONAL SMALL NODULE WHILE A LARGE SUSPICIOUS NODULE HAD ALREADY BEEN DETECTED BY THE RADIOLOGIST DID NOT IMPACT THE RADIOLOGIST’S FOLLOW-UP ADVICE. OF THE 3 ASSESSMENT LEVELS, FOLLOW-UP ADVICE IS CLINICALLY MOST RELEVANT. WHEN THE FOLLOW-UP ADVICE WAS ADJUSTED, IT WAS MOSTLY CHANGED TO A SHORTER FOLLOW-UP PERIOD (8/12, 67% ASSESSMENTS; EG, FROM CT AT 6-12 MONTHS TO CT AT 3-6 MONTHS). THIS FINDING INDICATES THAT, OWING TO THE AI RECOMMENDATIONS, RADIOLOGISTS TENDED TO BE MORE CAREFUL AND TOOK FEWER RISKS IN THEIR FOLLOW-UP ADVICE. FOR THIS STUDY, EARLIER FOLLOW-UP WAS APPROPRIATE AS ALL CT SCANS SHOWED CANCER CASES, BUT IN CLINICAL PRACTICE, IT CAN BE QUESTIONABLE WHETHER BEING MORE CAREFUL AND TAKING FEWER RISKS IN THE FOLLOW-UP ADVICE IS ALWAYS DESIRABLE BECAUSE IT MAY INCREASE THE HEALTH CARE COSTS. THEREFORE, IT IS OF GREAT IMPORTANCE TO STUDY THE COST-EFFECTIVENESS OF AI-CAD SYSTEMS.

SECONDARY FINDINGS

CONFIDENCE IN MALIGNANCY PREDICTION WAS SIGNIFICANTLY MORE FREQUENTLY CHANGED WHEN THE RADIOLOGIST DID NOT CHANGE THEIR NUMBER OF NODULES AFTER VIEWING THE AI RECOMMENDATIONS (MULTIMEDIA APPENDIX 6). THIS MIGHT BE CAUSED BY THE MALIGNANCY PREDICTION PROVIDED BY THE AI-CAD SYSTEM OF NODULES THAT THEY ALSO FOUND THEMSELVES. THE RADIOLOGIST MIGHT BECOME MORE CONVINCED WHETHER A CASE IS MALIGNANT OR BENIGN BASED ON THIS AI-CAD MALIGNANCY RECOMMENDATION.

THIS STUDY ALSO DEMONSTRATES THE IMPORTANCE OF APPLYING A USER-CENTERED DESIGN PROCESS TO ACHIEVE APPROPRIATE USE OF THE AI-CAD SYSTEM. THIS IS LACKING IN MANY STUDIES AND APPLICATIONS [46]. RADIOLOGISTS INDICATED IN THEIR FEEDBACK THAT THE PPV AND NPV WERE DIFFICULT TO INTERPRET. THEREFORE, DIFFERENT VISUALIZATIONS OF MODEL CONFIDENCE MIGHT BE MORE APPROPRIATE, SUCH AS USING ONLY BAR GRAPHS. FURTHERMORE, RADIOLOGISTS MENTIONED THAT SOME EXTRA FUNCTIONALITIES THAT RADIOLOGISTS USE IN CLINICAL PRACTICE FOR LUNG ASSESSMENT NEED TO BE IMPLEMENTED IN THE PROTOTYPE, SUCH AS MULTIPLANAR RECONSTRUCTION AND MAXIMUM INTENSITY PROJECTION, UNDERLINING THE NEED FOR TIGHT INTEGRATION OF AI INTO THE RADIOLOGIST ROUTINE WORKSTATIONS. IN ADDITION, THEY MENTIONED THAT DURING ONBOARDING, THEY WOULD LIKE TO RECEIVE MORE INFORMATION ON AI MODEL TRAINING AND VALIDATION, INCLUDING THE DATA SETS USED AND GROUND TRUTH DEFINITION, WHICH SHOULD THEREFORE BE ADDED TO THE ONBOARDING PROTOTYPE. THIS NEED IS IN LINE WITH THE FINDINGS OF CAI ET AL [31], WHO EXPLORED THE INFORMATION NEEDS FOR ONBOARDING FOR AI-CAD IN PATHOLOGY. ASHOORI AND WEISZ [43] MENTIONED THAT INFORMATION ON AI MODEL TRAINING AND TESTING IS IMPORTANT FOR RADIOLOGISTS’ TRUST IN AI-CAD SYSTEMS. RADIOLOGISTS’ FEEDBACK NEEDS TO BE INCORPORATED TOACHIEVE THE AI-CAD SYSTEM THAT FULLY MEETS RADIOLOGISTS’ NEEDS.

LIMITATIONS AND FUTURE PERSPECTIVES

THIS STUDY HAD SEVERAL LIMITATIONS. FIRST, THIS STUDY WAS NOT FULLY REPRESENTATIVE OF THE CLINICAL SITUATION. OWING TO TIME CONSTRAINTS, WE SPECIFICALLY ASKED THE RADIOLOGISTS NOT TO ASSESS THE ENTIRE CASE BUT TO FOCUS ON THE COMPONENT TASK OF LUNG NODULE ASSESSMENT. THEREFORE, RADIOLOGISTS WERE AWARE THAT LUNG NODULE ASSESSMENT WAS IMPORTANT, WHICH IS REPRESENTATIVE FOR CT SCANS ACQUIRED DUE TO PULMONARY COMPLAINTS BUT NOT FOR SCANS WITH INCIDENTAL LUNG NODULES. IN ADDITION, THIS STUDY EXCLUSIVELY INCLUDED SCANS OF CANCER CASES, WHICH DIFFERS FROM CLINICAL PRACTICE, IN WHICH SCANS MAY ALSO SHOW NO NODULES AND SOLELY BENIGN NODULES. HOWEVER, THE DATA SET WITH CANCER CASES WAS APPROPRIATE FOR OUR RESEARCH GOALS.

SECOND, IN THE CURRENT PROTOTYPE, THE EXPLAINABLE AI OUTPUT WAS SIMULATED POST HOC. THERE IS AN INCREASINGLY LOUDER CALL TO BUILD CAUSAL MODELS IN THE MEDICAL DOMAIN WHERE THE COST OF FAILURE IS HIGH, ALLOWING THE CLINICIAN TO VERIFY THE CAUSAL CHAIN OF EFFECTS OF CLINICALLY VALIDATED FEATURES ON THE MODEL PREDICTION. HOWEVER, SUCH INHERENTLY INTERPRETABLE MODELS ARE CURRENTLY THE EXCEPTION RATHER THAN MAINSTREAM PRACTICE [47]. IN THIS STUDY, WE FOCUSED ON THE CURRENT STATE OF MEDICAL PRACTICE, WHERE, IF AT ALL, MOST POST HOC EXPLAINABILITY TECHNIQUES ARE USED TO IMPROVE INTERPRETABILITY. IMPORTANTLY, POST HOC TECHNIQUES COME AT THE EXPENSE OF THE VALIDITY OF THE RELATIONSHIP BETWEEN POST HOC EXPLANATIONS AND MODEL PREDICTION. IN FACT, WHAT APPEARS TO AN END USER AS AN EXPLANATION MIGHT NOT CONVEY WHY THE BLACK BOX PREDICTED WHAT IT DID [48]. IN THIS STUDY, WE WERE INTERESTED IN THE EFFECT OF A WIDESPREAD APPROACH TO EXPLAIN USER TRUST AND DECISION-MAKING IN A MEDICAL CONTEXT.

https://ai.jmir.org/2024/1/e52211

(EN)
Simulating explainable AI output is very useful in the early stages of AI-CAD system development [33,34], having fully functioning AI models would further add to the realism of the test. Furthermore, it would be valuable if the algorithm can provide the extent to which each nodule characteristic contributed to malignancy prediction. In addition, PPV and NPV computed at the patient level were applied at the nodule level.

Third, this study included only 20 radiologists and 7 CT scans, which need to be scaled up to have sufficient power to be able to detect smaller effect sizes. In this pilot study, this limitation was accepted to make the test less time-consuming for the participating radiologists and to postpone larger samples after at least some evidence of larger effects in this context could be established. During case selection for this study, we aimed to collect a mix of relatively easy and more challenging cases, which worked well, considering the number of correct follow-up recommendations in Table 1. In a future large-scale study, it would be advisable to use a clinically representative data set to prevent the impact of selection bias. Testing on a larger scale is also required to analyze what radiologists do with FP findings and how these findings affect their trust in the AI-CAD. It is interesting to assess which types of FP findings are recognized by radiologists. Furthermore, it is useful to analyze whether changes in the number of observed nodules and in malignancy probability are correct based on a reference standard defined by expert radiologists and pathology. This is important because of automation bias, implying that radiologists rely too much on the AI recommendations, has to be prevented [40,49].

**Conclusions**

When clinical decision support systems are implemented, clinicians should receive careful onboarding that gives them a better understanding of the capabilities and limitations of the AI-CAD system. This understanding contributes to appropriate trust in the AI system, which is important when AI systems are used in clinical practice. Providing more AI output transparency, which enables clinicians to observe inconsistencies in the AI recommendations, can decrease clinicians’ trust in the AI-CAD system. AI recommendations frequently increased radiologists’ confidence in their assessments, even if they did not fully agree with these recommendations.

**Acknowledgments**

The members of the e/MTIC Oncology group are Fons van der Sommen, Joost Nederend, Misha D P Luyer, Mathias Funk, Jon R Pluyter, Igor Jacobs, Dimitrios Mavroeidis, Chris C P Snijders, Susan Hommerson, Lotte J S Ewals, Mark Ramaekers, Kasper van der Wulp, Christiaan G A Viviers, Terese A E Hellström, Nick H C Ruijs, Ning Fang and Victoria Bruno.

**Conflicts of Interest**

None declared.

**Multimedia Appendix 1**

Characteristics of computed tomography cases.
[DOCX File, 17 KB - ai_v3i1e52211_app1.docx]

**Multimedia Appendix 2**

Positive predictive value and negative predictive value definitions.
[DOCX File, 15 KB - ai_v3i1e52211_app2.docx]

**Multimedia Appendix 3**

Experimental conditions.
[DOCX File, 1239 KB - ai_v3i1e52211_app3.docx]

**Multimedia Appendix 4**

Forms for measuring trust.
[DOCX File, 24 KB - ai_v3i1e52211_app4.docx]

**Multimedia Appendix 5**

Mental model and psychological trust.
[DOCX File, 16 KB - ai_v3i1e52211_app5.docx]

**Multimedia Appendix 6**

Use of artificial intelligence recommendations.
[DOCX File, 22 KB - ai_v3i1e52211_app6.docx]

**References**

https://ai.jmir.org/2024/1/e52211


Abbreviations

- **AI**: artificial intelligence
- **AI-CAD**: artificial intelligence–based computer-aided detection or diagnosis
- **CT**: computed tomography
- **FP**: false-positive
- **NPV**: negative predictive value
- **PPV**: positive predictive value

©Lotte J S Ewals, Lynn J J Heesterbeek, Bin Yu, Kasper van der Wulp, Dimitrios Mavroeidis, Mathias Funk, Chris C P Snijders, Igor Jacobs, Joost Nederend, Jon R Pluyter, e/MTIC Oncology group. Originally published in JMIR AI (https://ai.jmir.org), 13.03.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR AI, is properly cited. The complete bibliographic information, a link to the original publication on https://www.ai.jmir.org/, as well as this copyright and license information must be included.
Research Letter

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

Paul Sebo¹, MSc, MD
University Institute for Primary Care, University of Geneva, Geneva, Switzerland

Corresponding Author:
Paul Sebo, MSc, MD
University Institute for Primary Care
University of Geneva
Rue Michel-Servet 1
Geneva, 1211
Switzerland
Phone: 41 223794390
Email: paulsebo@hotmail.com

(JMIR AI 2024;3:e53656) doi:10.2196/53656

KEYWORDS
accuracy; artificial intelligence; AI; ChatGPT; gender; gender detection tool; misclassification; name; performance; gender detection; gender detection tools; inequalities; language model; NamSor; Gender API; Switzerland; physicians; gender bias; disparities; gender disparities; gender gap

Introduction

Accurate determination of gender from names is vital for addressing gender-related disparities in medicine and promoting inclusivity. Gender detection tools (GDTs) offer efficient solutions, enabling large-scale demographic analysis [1-3] to improve data quality and inform targeted interventions. Indeed, they can process thousands of names simultaneously, saving time and resources. However, most of them charge for more than a certain number of requests per month. We recently compared the performance of 4 GDTs and showed that Gender API (Gender-API.com) and NamSor (NamSor Applied Onomastics) were accurate (misclassifications=1.5% and 2.0%, respectively; nonclassifications=0.3% and 0%, respectively) [4].

ChatGPT is a language model developed by OpenAI that is capable of generating human-like text and engaging in natural language conversations [5]. In medicine, ChatGPT can be employed for various purposes, such as answering patient queries and providing information on medical topics, making it a valuable resource for health care professionals and researchers seeking quick access to medical information and support in their work [6,7].

Given the increasing usefulness of GDTs in research, particularly for evaluating gender disparities in medicine, we assessed whether the performance of ChatGPT as a free GDT (version GPT-3.5) could approach that of Gender API and NamSor. We also compared ChatGPT-3.5 with the more advanced GPT-4 version. We hypothesized that ChatGPT, a versatile language model not specifically trained for gender analysis, could achieve gender detection performance comparable to specialized tools and that ChatGPT-4 would perform no better than ChatGPT-3.5.

Methods

Database Selection and Data Collection
The methods used in this study are the same as those used in our primary study, which compared the performance of 4 GDTs [4]. We used a database of 6131 physicians practicing in Switzerland, a multilingual and multicultural country with 36% of physicians of foreign origin [4]. The sample consisted of 3085 women (50.3%) and 3046 men (49.7%), with gender determined by self-identification. We used nationalize.io to determine the origin of physicians’ names (Table 1). A total of 88% of names were from French-, English-, Spanish-, Italian-, German-, or Portuguese-speaking countries or from another European country.

We asked ChatGPT-3.5 to determine the gender of 500 physicians at a time, after copying and pasting these lists of first and last names from the database. We ran the analysis twice and also examined ChatGPT-4 to check the “stability” of the responses [8]. The data were collected between September and November 2023.

We constructed a confusion matrix (Table 2): ff and mm correspond to correct classifications, mf and fm to misclassifications, and fu and mu to nonclassifications (ie, gender impossible to determine).
As in other studies [4,9], we calculated 4 performance metrics, namely “errorCoded” (the proportion of misclassifications and nonclassifications), “errorCodedWithoutNA” (the proportion of misclassifications), “naCoded” (the proportion of nonclassifications), and “errorGenderBias” (the direction of bias in gender determination). We used Cohen \( \kappa \) to assess interrater agreement.

Table 1. Estimated origin of physicians’ names (N=6131 physicians).

<table>
<thead>
<tr>
<th>Origin</th>
<th>Count(^a), n (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>French-speaking country</td>
<td>1679 (32.2)</td>
</tr>
<tr>
<td>English-speaking country</td>
<td>751 (14.4)</td>
</tr>
<tr>
<td>Spanish-speaking country</td>
<td>404 (7.7)</td>
</tr>
<tr>
<td>Asian country(^b)</td>
<td>344 (6.6)</td>
</tr>
<tr>
<td>Eastern European country</td>
<td>324 (6.2)</td>
</tr>
<tr>
<td>Italian-speaking country</td>
<td>288 (5.5)</td>
</tr>
<tr>
<td>Western European country(^b)</td>
<td>272 (5.2)</td>
</tr>
<tr>
<td>Arabic-speaking country</td>
<td>259 (5.0)</td>
</tr>
<tr>
<td>German-speaking country</td>
<td>259 (5.0)</td>
</tr>
<tr>
<td>Northern European country(^b)</td>
<td>220 (4.2)</td>
</tr>
<tr>
<td>Southern European country(^b)</td>
<td>217 (4.2)</td>
</tr>
<tr>
<td>Portuguese-speaking country</td>
<td>198 (3.8)</td>
</tr>
</tbody>
</table>

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

\(^b\)If not already classified in another group (eg, in the Arabic-speaking country group for some Asian countries).

Table 2. Confusion matrix showing the 6 possible classification outcomes.

<table>
<thead>
<tr>
<th></th>
<th>Female (predicted)</th>
<th>Male (predicted)</th>
<th>Unknown (predicted)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Female (actual)</td>
<td>ff</td>
<td>fm</td>
<td>fu</td>
</tr>
<tr>
<td>Male (actual)</td>
<td>mf</td>
<td>mm</td>
<td>mu</td>
</tr>
</tbody>
</table>

Ethical Considerations
Since this study did not involve the collection of personal health–related data, it did not require ethical review per current Swiss law.

Results
Performance metrics showed high accuracy for ChatGPT-3.5 and ChatGPT-4 in both the first and second rounds (Table 3).

The number of misclassifications was low (proportion\( \leq 1.5\% \)) and there were no “nonclassifications.” As shown in Table 3, interrater agreement between the first and second rounds (for ChatGPT-3.5 and ChatGPT-4) and between ChatGPT-3.5 and ChatGPT-4 (for the first round) was “almost perfect” (\( \kappa > 0.97 \), all \( P < .001 \)).
Table 3. Confusion matrix and performance metrics for ChatGPT-3.5 and ChatGPT-4 (N=6131 physicians).

<table>
<thead>
<tr>
<th></th>
<th>Classified as women, n (%)</th>
<th>Classified as men, n (%)</th>
<th>Unclassified, n (%)</th>
<th>Interrater agreement&lt;sup&gt;a&lt;/sup&gt;</th>
<th>Cohen κ (95% CI)</th>
<th>P value</th>
</tr>
</thead>
<tbody>
<tr>
<td>ChatGPT-3.5</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>First round&lt;sup&gt;c&lt;/sup&gt;</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Female physicians (n=3085)</td>
<td>3028 (98.2)</td>
<td>57 (1.8)</td>
<td>0 (0)</td>
<td></td>
<td>0.9817 (0.9770-0.9865)&lt;sup&gt;b&lt;/sup&gt;</td>
<td>&lt;.001</td>
</tr>
<tr>
<td>Male physicians (n=3046)</td>
<td>18 (0.6)</td>
<td>3028 (99.4)</td>
<td>0 (0)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Second round&lt;sup&gt;d&lt;/sup&gt;</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Female physicians (n=3085)</td>
<td>3030 (98.2)</td>
<td>55 (1.8)</td>
<td>0 (0)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Male physicians (n=3046)</td>
<td>28 (0.9)</td>
<td>3018 (99.1)</td>
<td>0 (0)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ChatGPT-4</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>First round&lt;sup&gt;e&lt;/sup&gt;</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Female physicians (n=3085)</td>
<td>3020 (97.9)</td>
<td>65 (2.1)</td>
<td>0 (0)</td>
<td></td>
<td>0.9958 (0.9935-0.9981)&lt;sup&gt;b&lt;/sup&gt;</td>
<td>&lt;.001</td>
</tr>
<tr>
<td>Male physicians (n=3046)</td>
<td>27 (0.9)</td>
<td>3019 (99.1)</td>
<td>0 (0)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Second round&lt;sup&gt;f&lt;/sup&gt;</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Female physicians (n=3085)</td>
<td>3020 (97.9)</td>
<td>65 (2.1)</td>
<td>0 (0)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Male physicians (n=3046)</td>
<td>26 (0.9)</td>
<td>3020 (99.1)</td>
<td>0 (0)</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

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

<sup>b</sup>Interrater agreement between the first and second rounds for each version.

<sup>c</sup>Performance metrics: errorCoded=0.01223, errorCodedWithoutNA=0.01223, naCoded=0, and errorGenderBias=–0.00636.

<sup>d</sup>Performance metrics: errorCoded=0.01354, errorCodedWithoutNA=0.01354, naCoded=0, and errorGenderBias=–0.00440.

<sup>e</sup>Performance metrics: errorCoded=0.01501, errorCodedWithoutNA=0.01501, naCoded=0, and errorGenderBias=–0.00620.

<sup>f</sup>Performance metrics: errorCoded=0.01484, errorCodedWithoutNA=0.01484, naCoded=0, and errorGenderBias=–0.00636.

**Discussion**

We used ChatGPT to determine the gender of 6131 physicians practicing in Switzerland and found that the proportion of misclassifications was ≤1.5% for both versions. There were no nonclassifications and gender bias was negligible. Interrater agreement between ChatGPT-3.5 and ChatGPT-4 was “almost perfect.”

These results are relatively similar to those found in our primary study for Gender API and NamSor (errorCoded=0.0181 and 0.0202, errorCodedWithoutNA=0.0147 and 0.0202, naCoded=0.0034 and 0, errorGenderBias=–0.0072 and 0.0026) [4]. They are slightly better than those of another study published in 2018, which compared 5 GDTs, including Gender API and NamSor [9]. These results suggest that ChatGPT can accurately determine the gender of individuals using their first and last names. The disadvantage of ChatGPT compared to Gender API and NamSor is that the database cannot be uploaded directly into ChatGPT (eg, as an Excel or CSV file).

Both ChatGPT-3.5 and ChatGPT-4 exhibit high accuracy in gender detection, with no significant superiority observed in ChatGPT-4 over ChatGPT-3.5. This underscores the robustness of ChatGPT in gender prediction across different versions. Our short study has 2 main limitations. Given the estimated origin of physicians’ names, the results of the study can probably be generalized to most Western countries but not necessarily to Asian or Middle Eastern countries. GDTs are often less accurate with names from these countries [9,10]. In addition, GDTs oversimplify the concept of gender by dichotomizing individuals into male or female.

**Data Availability**

The data associated with this article are available in the Open Science Framework [11].

**Conflicts of Interest**

None declared.

**References**

https://ai.jmir.org/2024/1/e53656


11. What is the performance of ChatGPT in determining the gender of individuals based on their first and last names? Open Science Framework. 2023 Sep 27. URL: https://osf.io/6nzd4/ [accessed 2024-03-08]

**Abbreviations**

GDT: gender detection tool

---

**Edited by K El Emam, B Malin; submitted 14.10.23; peer-reviewed by ZA Teel, A Shamsi, L Zhu; comments to author 21.11.23; revised version received 26.11.23; accepted 02.03.24; published 13.03.24.**

**Please cite as:**

Sebo P

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

JMIR AI 2024;3:e53656

URL: https://ai.jmir.org/2024/1/e53656
doi:10.2196/53656
PMID:

©Paul Sebo. Originally published in JMIR AI (https://ai.jmir.org), 13.03.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR AI, is properly cited. The complete bibliographic information, a link to the original publication on https://www.ai.jmir.org/, as well as this copyright and license information must be included.
Leveraging Machine Learning to Develop Digital Engagement Phenotypes of Users in a Digital Diabetes Prevention Program: Evaluation Study

Danissa V Rodriguez¹, PhD; Ji Chen¹, PhD; Ratnalekha V N Viswanadham¹, PhD; Katharine Lawrence¹,², MPH, MD; Devin Mann¹,², MS, MD

¹New York University Grosman School of Medicine, New York, NY, United States
²New York University Langone Health, New York, NY, United States

Abstract

Background: Digital diabetes prevention programs (dDPPs) are effective “digital prescriptions” but have high attrition rates and program noncompletion. To address this, we developed a personalized automatic messaging system (PAMS) that leverages SMS text messaging and data integration into clinical workflows to increase dDPP engagement via enhanced patient-provider communication. Preliminary data showed positive results. However, further investigation is needed to determine how to optimize the tailoring of support technology such as PAMS based on a user’s preferences to boost their dDPP engagement.

Objective: This study evaluates leveraging machine learning (ML) to develop digital engagement phenotypes of dDPP users and assess ML’s accuracy in predicting engagement with dDPP activities. This research will be used in a PAMS optimization process to improve PAMS personalization by incorporating engagement prediction and digital phenotyping. This study aims (1) to prove the feasibility of using dDPP user-collected data to build an ML model that predicts engagement and contributes to identifying digital engagement phenotypes, (2) to describe methods for developing ML models with dDPP data sets and present preliminary results, and (3) to present preliminary data on user profiling based on ML model outputs.

Methods: Using the gradient-boosted forest model, we predicted engagement in 4 dDPP individual activities (physical activity, lessons, social activity, and weigh-ins) and general activity (engagement in any activity) based on previous short- and long-term activity in the app. The area under the receiver operating characteristic curve, the area under the precision-recall curve, and the Brier score metrics determined the performance of the model. Shapley values reflected the feature importance of the models and determined what variables informed user profiling through latent profile analysis.

Results: We developed 2 models using weekly and daily DPP data sets (328,821 and 704,242 records, respectively), which yielded predictive accuracies above 90%. Although both models were highly accurate, the daily model better fitted our research plan because it predicted daily changes in individual activities, which was crucial for creating the “digital phenotypes.” To better understand the variables contributing to the model predictor, we calculated the Shapley values for both models to identify the features with the highest contribution to model fit; engagement with any activity in the dDPP in the last 7 days had the most predictive power. We profiled users with latent profile analysis after 2 weeks of engagement (Bayesian information criterion=−3222.46) with the dDPP and identified 6 profiles of users, including those with high engagement, minimal engagement, and attrition.

Conclusions: Preliminary results demonstrate that applying ML methods with predicting power is an acceptable mechanism to tailor and optimize messaging interventions to support patient engagement and adherence to digital prescriptions. The results enable future optimization of our existing messaging platform and expansion of this methodology to other clinical domains.

Trial Registration: ClinicalTrials.gov NCT04773834; https://www.clinicaltrials.gov/ct2/show/NCT04773834

International Registered Report Identifier (IRRID): RR2-10.2196/26750
machine learning; digital health; diabetes; mobile health; messaging platforms; user engagement; patient behavior; digital diabetes prevention programs; digital phenotypes; digital prescription; users; prevention; evaluation study; communication; support; engagement; phenotypes; digital health intervention; chronic disease management

Introduction

Over 80 million US adults have prediabetes, a metabolic condition that places individuals at risk for progression to type 2 diabetes and its related complications [1]. Evidence-based strategies for diabetes prevention have primarily focused on nonpharmacologic interventions such as diabetes prevention programs (DPPs), which are comprehensive behavior change curricula concentrating on physical activity and dietary modification. Such programs can be as effective as medication in preventing the progression of diabetes in at-risk populations [2]. Increasingly, DPP behavioral curricula have been adapted to digital platforms (digital DPPs [dDPPs]), which have demonstrated comparable effectiveness in achieving weight loss, hemoglobin A1c reduction, and other critical diabetes-related health outcomes while offering improvements in accessibility, convenience, and personalization [3]. Yet, limited patient engagement with digital interventions presents a significant barrier to translating evidence-based digital behavioral interventions such as the dDPP into pragmatic, scalable solutions [4-8].

To address this critical patient engagement issue, various technologies and interventions have been developed to provide targeted support to patients using digital health apps to improve engagement and sustained use [9]. Potential solutions include mobile-based feedback and reminder tools, app-based coaching, social networking, and gamification. More recent strategies have also leveraged machine learning (ML) and big data analytics to deploy more advanced tools, such as engagement algorithms and artificial intelligence (AI)–driven chatbots. ML solutions can provide (1) more nuanced patient segmentation or phenotyping; (2) more precise, tailored interventions, with enhanced ability to respond dynamically to changes in individual trends; and (3) improved resource alignment by intervention implementers, as automated processes (eg, chatbots) can free up human capital for more appropriate tasks [10]. Moreover, AI-driven chatbots (AI chatbots), conversational agents that mimic human interaction through written, oral, and visual communication channels with a user [1,2], have demonstrated efficacy in health-behavior change interventions among a large and diverse population [3-6,11-13].

Prior work from this team involved developing a personalized automatic messaging system (PAMS) that leveraged an evidence-based engagement algorithm to deliver tailored behavior change theory–supported SMS text messaging to support users engaging with a commercial app-based dDPP. The study returned promising results compared with average users, demonstrating engagement in various dDPP features (eg, weight tracking and physical activity logins) [12]. To expand on the previous investigation, improved features of the next generation of PAMS include an ML-based patient engagement prediction algorithm to identify dDPP digital engagement phenotypes and to guide and further personalize the messaging intervention. This paper describes the ML model designed to predict characteristics and behavioral patterns of dDPP user types (eg, those highly engaged with exercise but not uploading the meals or those messaging their coach but not participating in weigh-ins) based on their activity patterns within a dDPP app, with a particular focus on motivating users at risk for low engagement and nonengagement with the dDPP (ie, patient digital engagement phenotypes).

Methods

Overview

The logic diagram in Figure 1 illustrates, from left to right, the overall framework for optimizing patient engagement with a dDPP [14]. In this study, we completed 2 activities (developing, validating, and testing ML models and studying model outputs with latent profile analysis [LPA]) and identified future activities toward optimization. The drivers behind this optimization initiative stem from low levels of patient engagement with dDPPs and other wellness-based mobile apps. We used the daily and weekly data sets provided by the dDPP vendor (inputs) to develop, validate, and test an ML model for each data set (first activity). On the basis of the performance metrics from the daily and weekly models, we identified the highest contributing feature for each model using Shapley values (first outputs). These features were fed into the LPA (second activity) to determine the number of participant usage profiles (second outputs). The goodness of fit derived from the LPA validated the phenotypes formed from the LPA (direct outcome). This integration of ML and statistical learning processes would inform how we identify digital engagement phenotypes for the dDPP study set (in the dashed red box) and, therefore, design content for a more personalized messaging platform (second direct outcome). Ultimately, the desired long-term outcomes of the profiling process are increased patient engagement with the dDPP and a reduction in clinical outcomes related to hemoglobin A1c and weight (indirect outcomes). The process rests on the assumptions that the dDPP data accurately reflect digital behavioral patterns and that people from the vendor-provided data are representative of people in the study data set.
**Figure 1.** Logic diagram of the research methodology to integrate machine learning (ML) into participant profiling, including the input data sets; the methods applied to the data sets; and the intermediary, direct, and indirect outputs. dDPP: digital diabetes prevention program.

**Participants**

Study participants were users with prediabetes who enrolled in a commercial dDPP app (our dDPP research vendor), including nonpatient (“vendor”) users and institution-based patients (“study” participants of this dDPP intervention) [11]. Eligible participants are at least 18 years old, have a BMI of at least 25 kg/m² (22 kg/m² if self-identified as Asian), have a diagnosis of prediabetes (either by International Classification of Diseases, Tenth Revision code, problem list, or a hemoglobin A1c level of 5.7%-6.4% in the last 12 months), and are deemed safe to engage in light physical exercise and weight loss by their primary care physician. For institutional study participants enrolled in the current clinical trial of this dDPP intervention, patients are excluded if they have a prior diagnosis of diabetes, have any end-stage illness with a prognosis within 6 months, are non-English speakers (as the dDPP program is currently only available in English), or are unable to send or receive SMS text messages [4]. Recruited patients were identified via electronic health record review and contacted through multichannel methods (eg, patient portal, email, in-clinic recruitment, and clinician referral).

**The Data**

**Data Sourcing**

Data for the evaluation were sourced from a commercial dDPP vendor and a patient cohort of an academic health center. We used 2 deidentified data sets (weekly and daily data) of eligible retail users for the initial training, validation, and testing of the ML models. These data sets aggregate and present user information on a weekly or daily basis and capture all features recorded by the dDPP app, including per user or patient: meals logged, steps logged, exercises logged, messages shared with the dDPP coach and other dDPP patients using the app, app log-ins, and the number of dDPP articles read. These activities were the same as those used for generating the adherence algorithm in our previous research. In addition to the vendor-provided data sets, for a later testing phase, we use an existing data set of data collected from dDPP patients who are part of this dDPP study and exposed to the PAMS intervention.

**Weekly dDPP Vendor Data Set**

Data include detailed information about all the features collected for our dDPP app partners, such as meals logged, steps logged, exercises logged, messages shared with the dDPP coach and other dDPP patients using the app, app log-ins, and the number of dDPP articles read during each week. All users have more than 5 weeks of engagement records, and we used only 1 year’s worth of dDPP engagement data per user.

**Weekly dDPP Institutional Study Data Set**

The 2 data sets (weekly dDPP vendor data set and weekly dDPP study data set) have the same data structure. The same data fields are collected for commercial users and the dDPP patients, but the only difference is on the behavioral level because the patients’ data are potentially affected by the message intervention (PAMS). All data were used for the validation of the weekly ML model.

**Daily dDPP Vendor Data Set**

In addition to the activity records in the weekly data, we had access within the daily data set to calorie consumption data, meal logs, and color codes assigned to each food item as reported by the users. Users with less than 7 days of engagement records were excluded from the cohort, and we used only 1 year’s worth of dDPP engagement data per user.

**Outcomes**

First, we built binary classification ML models to predict whether a participant will engage in the next week or the next day with the dDPP based on their previous short- and long-term activity in the app. For the weekly model, we used the vendor data set to train and validate retrospectively to predict general activity (engagement in any activity). We prospectively validated the weekly model using the institutional study data.
set. For the daily models, we predicted 5 outcomes: general activity, physical activities (steps and exercises recorded on the app), in-app lessons (article reading), social activities (group posts and coach messages in the app), and weigh-ins in the app. Second, we identified the variables from the daily overall activity model of the vendor’s participants that provide the most predictive power for engagement. Third, we evaluated whether these predictive variables could generate profiles of a participant’s behavior that can be targeted with motivational messaging.

Predictors
We built model predictors from users’ demographic data and collected in-app activities. These activities include steps taken, exercises, meal logs, weigh-in records, in-app messaging and group activities, and in-app article reading. For the weekly data set, short-term activity profiles were built from the week before the evaluation week and up to 4 weeks before the evaluation week. Long-term activity profiles were summarized and constructed from the first week of program enrollment up to the evaluation week. Short-term activity profiles were built from the day before and within 7 days before the evaluation day for the daily data set. Similarly, long-term activity profiles were summarized and constructed from the first day of program enrollment up to the evaluation day. The day of the week and national holidays were also captured as predictors. In total, 43 predictors were used to build weekly models, and 49 predictors were used to build daily models.

Sample Size
The sample sizes for user weekly and daily data sets directly from the dDPP vendor were determined by the convenience of the dDPP vendor and assumed to be representative of the academic health center’s study sample. The study sample size was determined by the number of participants already recruited and actively involved in the original dDPP study as of December 2021 [4].

Missing Data
Because this paper aims to predict participant engagement with the dDPP, missing data among in-app activities were treated as a participant not engaging in either overall activity (ie, no observations for a particular day or a week for any activity) or specific within-dDPP activities (eg, a participant not recording meals or reading any articles). Missing participant weight was logged as a participant not weighing themselves for the dDPP, and we ignored the magnitude of weight due to individual non-dDPP factors contributing to weight outcomes. No participant had a missing age due to age being a requirement for enrollment into the dDPP. Participants who did not record their ideal body weight at the beginning of dDPP engagement had this observation recorded as a 0, as the lack of goal recording for weight could have clinical implications (eg, weight is not the primary utilization goal for the participant, or the participant is not comfortable with setting a weight goal). No participant had a missing initial BMI recorded. One participant was missing gender identification, so their observations were removed from the data set.

Statistical Analysis Methods
Data Split
All data sets were split into a 70% training set, a 15% validation set, and a 15% test set based on users. Observations of any user only existed in 1 set to prevent potential data leak and unintended bias.

Gradient-Boosted Forest Algorithm
We use the gradient-boosted forest algorithm, a robust regression tree approach that includes multiple simple decision trees to iteratively refine the performance of the model by minimizing the difference between the expected and expert-labeled outcomes [15,16]. Forest-based algorithms provide 2 fundamental benefits. First, they allow for nonlinear interactions between covariates to impact the prediction of the dependent variable, as opposed to using a Least Absolute Shrinkage and Selection Operator (LASSO) or a ridge regression model. Second, forest-based algorithms do not require a priori function structure to define the relationship between the covariates and the outcome. For example, we do not need to theoretically assume whether a particular engagement type (eg, steps) interacts with another type (eg, exercise logging). We used gradient boosting to allow for prediction despite the sparsity of the data, as users may engage with one activity but not others on a given day or have no activity (ie, all observations as 0). The values defining engagement included binary predictors, large integers (eg, calories and steps), and values between 0 and 1 (eg, the portion of engagement throughout enrollment). These models aimed to identify that the sub-behaviors that create the most predictive power for engagement with the dDPP were trained with $\eta=0.1$ for 1000 rounds with early stopping.

Metrics
The area under the receiver operating characteristic curve (AUROC), the area under the precision-recall curve (AUPRC), and the Brier score statistics measured the performance of the model. To estimate the CIs of the evaluation metrics for the ML models, we performed bootstrapping with 200 iterations on the test set. In each iteration, a random sample of the test set, with replacement, was drawn with the same size as the original test set. The ML model was then evaluated on this bootstrapped sample, and the performance metrics mentioned above were recorded. The process was repeated for 200 iterations, resulting in a distribution of performance metrics from which the 95% CIs were calculated, providing a robust estimate of the performance and variability of the model. In addition, Shapley values were calculated to reflect the feature importance of each model.

Engagement Profiling
A person-centered approach to messaging can help motivate individuals to complete goal-oriented behaviors like engagement with a lifestyle management app [17]. This approach involves (1) tailoring delivery based on the person’s behavior profile within the app and (2) focusing messaging on targetable behaviors to motivate users to complete small, manageable actions toward their goal (ie, the goal gradient hypothesis in decision-making) [18]. We performed an LPA on the participants in the daily data set to determine the subgroups of

https://ai.jmir.org/2024/1/e47122

(parge number not for citation purposes)
participants’ behaviors. LPA identifies latent clusters of individuals based on continuous variables [19]. The contributions of multiple variables (ie, the facets that explain the unobserved profile of a user) contribute to the outcome experienced by a user. We used the covariates with the highest global mean Shapley values from the gradient-boosted forest model for the LPA for 2 reasons. First, these variables offer the most explanatory power behind the probability of engagement with the dDPP, allowing us not to assume a priori the behaviors that contribute to the usage of the dDPP. Second, profiling users of a digital app such as this dDPP can be more complicated than traditional approaches to consumer profiling, given the interaction between a user’s health and app engagement. To determine the minimum usage data after enrollment into a dDPP to start profiling participants, we conducted LPAs after 2 weeks and iteratively added days until 3 weeks of engagement. We used the profiles from the timestamp with the lowest Bayesian information criterion (BIC), the established goodness-of-fit metric for LPA. We used the mclust package in RStudio (version 2022.12.0+353; Posit Software, PBC) to run the LPAs [20].

Development Versus Validation

We validated the weekly model prospectively using the weekly dDPP study data set. Detailed information about this data set is under the subsection “Participants” [15,16].

Ethical Considerations

In this DPP research, ethical standards and the protection of human participants are emphasized. The study is committed to adhering to regulations outlined in 45 CFR Part 46, ensuring the rights and welfare of participants. The NYU Langone Health institutional review board (IRB) played a crucial role in reviewing and approving the research, informed consent forms, and recruitment materials before participant enrollment (i20-01548). The informed consent process is described as an ongoing dialogue, emphasizing clear communication, comprehension, and the right to withdraw without adverse consequences. The consent forms, including verbal consent and a key information sheet, were submitted to the IRB for approval. Confidentiality measures are robust, complying with the Health Insurance Portability and Accountability Act (HIPAA), and a Certificate of Confidentiality from the National Institutes of Health was obtained. Data security is maintained through password protection, and research data are stored securely. The research emphasizes that stored data will only be used for this study, with no plans for future use in subsequent research. Overall, the research underscores the importance of ethical conduct, participant consent, and stringent confidentiality measures in the research process.

Moreover, the research underscores the importance of ethical conduct, rigorous IRB oversight, and robust confidentiality measures to safeguard the rights and well-being of study participants. Additionally, it highlights the meticulous documentation of the informed consent process and the secure handling of research data, ensuring compliance with regulations and promoting participant trust and privacy.

Results

Participants

Table 1 details the descriptive statistics for the 3 preprocessed data sets, including weekly and daily data for the dDPP user (dDPP vendor data sets) and the weekly data for the dDPP patients (dDPP study data). For the vendor-provided data sets, users engage with the app 54.2% (208,142/384,025) of the times in the weekly data compared with 38.9% (274,200/704,242) of the times in the daily data. The average engagement within individual activities is similar. “Steps taken” had the highest percentage of all activities in both data sets. For study data, the engagement percentage was higher (92.1%, 1253/1361), which could be attributed to the effects of PAMS messages.
Table 1. Descriptive statistics of users (N=12,262).

<table>
<thead>
<tr>
<th>Characteristic</th>
<th>Weekly dDPP(^a) vendor data (dDPP vendor users, n=10,053)</th>
<th>Weekly dDPP study data (dDPP study patients, n=50)</th>
<th>Daily dDPP vendor data (dDPP vendor users, n=2159)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Program length</td>
<td>38.2 weeks</td>
<td>27.22 weeks</td>
<td>326.2 days</td>
</tr>
<tr>
<td>Age (years) mean (SD)</td>
<td>47.6 (11.4)</td>
<td>N/A(^b)</td>
<td>N/A</td>
</tr>
<tr>
<td><strong>Sex, n (%)</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Male</td>
<td>1267 (12.6)</td>
<td>N/A</td>
<td>N/A</td>
</tr>
<tr>
<td>Female</td>
<td>8786 (87.4)</td>
<td>N/A</td>
<td>N/A</td>
</tr>
<tr>
<td>Engagement of any activity, n/N (%)</td>
<td>208,142/384,025 (54.2)</td>
<td>1253/1361 (92.1)</td>
<td>274,200/704,242 (38.9)</td>
</tr>
<tr>
<td>Engagement of steps taken, n/N (%)</td>
<td>208,142/384,025 (54.2)</td>
<td>1086/1361 (79.8)</td>
<td>244,823/704,242 (34.8)</td>
</tr>
<tr>
<td>Engagement of exercises, n/N (%)</td>
<td>77,957/384,025 (20.3)</td>
<td>349/1361 (25.6)</td>
<td>49,683/704,242 (7.1)</td>
</tr>
<tr>
<td>Engagement of meals logged, n/N (%)</td>
<td>137,865/384,025 (35.9)</td>
<td>924/1361 (67.9)</td>
<td>100,449/704,242 (14.3)</td>
</tr>
<tr>
<td>Engagement of weigh-ins, n/N (%)</td>
<td>137,481/384,025 (35.8)</td>
<td>523/1361 (38.4)</td>
<td>71,596/704,242 (10.2)</td>
</tr>
<tr>
<td>Engagement of article reading, n/N</td>
<td>118,280/384,025 (30.8)</td>
<td>573/1361 (42.1)</td>
<td>79,272/704,242 (11.2)</td>
</tr>
<tr>
<td>Engagement of group posts, n/N (%)</td>
<td>24,578/384,025 (6.4)</td>
<td>100/1361 (7.3)</td>
<td>45,113/704,242 (6.4)</td>
</tr>
</tbody>
</table>

\(^a\) dDPP: digital diabetes prevention program.
\(^b\) N/A: not applicable.

**Weekly Model (For Any Activity) Development and Performance**

We trained and tested the model to predict “any activity” (ie, the probability of the subsequent interaction with the dDPP based on whether the user interacted with any of the features of the dDPP app, such as exercise, meal, and weigh-ins) on the weekly dDPP vendor data set. The weekly model reported an AUROC of 0.97 (95% CI 0.97-0.97), an AUPRC of 0.98 (95% CI 0.98-0.98), and a Brier score of 0.061 (95% CI 0.060-0.063) in the test set (Figure 2). Because we also aimed to identify how individual variables contribute to predictions by the model, we calculated the Shapley value, which is the average marginal contribution of a variable to a model across the different combinations of including the variable in the model (eg, nonlinear contributions and splitting a forest into different branches with the variable). The Shapley value method has become the preferred technique for feature attribution in ML models, thanks to its robust and reliable performance [21].

**Figure 2.** AUROC (left) and AUPRC (right) performance metrics of the “any activity” weekly model in the test set of the weekly vendor data set (58,210 engagement records). The calibration plot shows that the model is well calibrated. AUPRC: area under the precision-recall curve; AUROC: area under the receiver operating characteristic curve.
Figure 3 displays the distribution of the 10 covariates with the highest calculated global mean Shapley value (i.e., which variables have the strongest predictive power, regardless of negative or positive impact, on the user’s engagement with the dDPP). A higher magnitude of the Shapley value (i.e., further from 0) indicates the strength of the variable in the model to predict a user’s engagement with the dDPP. A positive Shapley value indicates that the user is more likely to engage with the dDPP because of the variable (i.e., a positive predictor). A negative Shapley value suggests that the patient is less likely to engage with the dDPP due to the variable (i.e., a negative predictor). More purple values indicate a higher mean for the covariate of the individual (e.g., a more purple “exercise frequency” dot indicates that the user logged for nonstep physical activity more than other users did). The covariates with the most contribution to model prediction were those of short-term behaviors.

Figure 3. Shapley values of top 10 features in the “any activity weekly model.” Each dot on the plot represents an engagement record and is colored according to the value of the corresponding feature from high (purple) to low (yellow). Features are ranked in descending order from top to bottom on the y-axis (i.e., variables with the highest contribution to the model are on the top), with global mean Shapley values of each feature annotated next to them.

We tested our model using the weekly dDPP institutional study data set (prospective clinical data). The model achieved an AUROC of 0.92 (95% CI 0.89-0.94), an AUPRC of 0.99 (95% CI 0.99-0.99; Figure 4), and a Brier score of 0.072 (95% CI 0.063-0.081), suggesting high predictive power and operational potential for refining PAMS using this method. After analyzing the weekly dDPP study data set, we detected that this data set would be imbalanced because the prediction of the subsequent week’s activity would be based on whether a user engaged with any app activity, rather than a particular activity, within the dDPP, seen by the 92.1% engagement ratio, and the sample size was too low to yield unbiased testing results. Regardless of the limitation of the research data set, this analysis was proper in confirming the effectiveness of the weekly model.
Daily Model (for Any Activity) Development and Performance

We expanded a proportion of the weekly data set into a daily (more detailed) format and trained 5 new models. Figure 5 illustrates the ML model fit in the test set of the daily data set. Figure 6 displays the distribution of the covariates with the strongest predictive power (ie, the highest global mean Shapley value). Like the weekly model, engagement with any activity in the dDPP in the last 7 days had the most predictive power (a global mean Shapley value of 2.638). However, in contrast to the weekly model, features associated with long-term activity also had strong predictive power in the model.
Although the daily model for “any activity” returned a high AUROC and AUPRC, we aimed to generate predictions on each specific activity to inform our user profiling (digital engagement phenotypes) and consequently elevate the message personalization. Therefore, we developed 4 ML models, focusing on daily engagement with each key type of activity for a dDPP (physical activity, lessons, social activity, and weigh-ins). Table 2 displays the model fits for each of these “submodels.” For each activity, the model indicates highly predictive behavioral patterns among users. The “physical activity” and “social activity” daily models had higher AUROC performance with slightly lower AUPRC than the other daily models. All daily models show higher levels of calibration (a highest Brier score of 0.051) than the weekly model (a Brier score of 0.061).

Table 2. Performance metrics of each daily activity model in the test set.

<table>
<thead>
<tr>
<th>Model fit metrics</th>
<th>Any app activity</th>
<th>Physical activity (exercises and steps)</th>
<th>Lessons (article reading)</th>
<th>Social activity (group posts and coach messages)</th>
<th>Weigh-ins</th>
</tr>
</thead>
<tbody>
<tr>
<td>AUROC a (95% CI)</td>
<td>0.99 (0.99-0.99)</td>
<td>0.98 (0.98-0.98)</td>
<td>0.99 (0.99-0.99)</td>
<td>0.98 (0.98-0.98)</td>
<td>0.94 (0.94-0.94)</td>
</tr>
<tr>
<td>AUPRC b (95% CI)</td>
<td>0.98 (0.98-0.98)</td>
<td>0.74 (0.72-0.75)</td>
<td>0.91 (0.91-0.92)</td>
<td>0.74 (0.73-0.75)</td>
<td>0.65 (0.63-0.66)</td>
</tr>
<tr>
<td>Brier score (95% CI)</td>
<td>0.037 (0.036-0.038)</td>
<td>0.025 (0.025-0.026)</td>
<td>0.027 (0.026-0.028)</td>
<td>0.02 (0.023-0.024)</td>
<td>0.051 (0.050-0.052)</td>
</tr>
</tbody>
</table>

aAUROC: area under the receiver operating characteristic curve.
bAUPRC: area under the precision-recall curve.

Engagement Profiling Development and Performance

We profiled participants with their daily engagement data using LPA after 2 weeks of dDPP enrollment. To determine the optimal time to start profiling participants, we iteratively added 1 day of engagement and created profiles until 3 weeks after their enrollment in the dDPP. After 2 weeks of daily engagement data, profiling participants had the strongest LPA model fit (BIC=−3222.46), followed by the model fit from profiling with 3 weeks of data (BIC=−2903.19). The LPA model fits for 15 to 20 days of engagement were significantly worse (ie, higher...
BIC values) and, therefore, are not reported. The best-performing LPA model was ellipsoidal (there is some correlation between variables), had equal volume (the variances are equal across identified profiles), and had variable distributions between profiles (ie, the number of people per profile vary), and consisted of 6 profiles. Table 3 reports the mean engagement for each variable within and across the profiles of participants.

### Table 3. Mean engagement by profile and across profiles for key engagement variables.

<table>
<thead>
<tr>
<th>Key engagement variables</th>
<th>Subbehavior variable mean (SE)</th>
<th>Mean engagement across profiles (SD)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Profile 1 (n=16)</td>
<td>Profile 2 (n=91)</td>
</tr>
<tr>
<td>Any activity rate (last 7 days)</td>
<td>0.969 (0.085)</td>
<td>0.992 (0.045)</td>
</tr>
<tr>
<td>Long-term activity rate</td>
<td>0.942 (0.105)</td>
<td>0.998 (0.013)</td>
</tr>
<tr>
<td>Steps taken rate (last 7 days)</td>
<td>2572 (4661)</td>
<td>3909 (3968)</td>
</tr>
<tr>
<td>Long-term weight-in rate</td>
<td>0.507 (0.310)</td>
<td>0.139 (0.139)</td>
</tr>
<tr>
<td>Recent meal rate</td>
<td>0.906 (0.256)</td>
<td>0.397 (0.416)</td>
</tr>
<tr>
<td>Long-term step rate</td>
<td>0.438 (0.345)</td>
<td>0.998 (0.013)</td>
</tr>
<tr>
<td>Long-term meal log rate</td>
<td>0.856 (0.249)</td>
<td>0.463 (0.339)</td>
</tr>
<tr>
<td>Article reading rate (last 7 days)</td>
<td>2.01 (1.549)</td>
<td>0.278 (0.704)</td>
</tr>
</tbody>
</table>

The LPA identified attrition (users in profile 5 who showed consistently low engagement across variables) and behaviors that show points of continued engagement for users. Users in profile 6, for example, had a close-to-average engagement with the dDPP from weigh-ins with the app and logging steps, which are behaviors that require one-time interactions with the dDPP, given Bluetooth connections between smart devices and the dDPP. In contrast, users in profile 3 were highly engaged, as they consistently engaged more than the average user. Messaging to users in profile 3 should, therefore, differ from messaging to users in profile 5, given the differences in their efforts toward the dDPP. Users in profile 4 had a lower-than-average engagement with the dDPP but showed the highest engagement with the learning materials across all users. Clusters 1 and 2 showed similarly high short- and long-term engagements but differed in engagement with the dDPP. Users in profile 1 read more educational materials provided in the dDPP, whereas users in profile 2 were more consistent in taking steps.

### Discussion

**Summary**

The literature suggests the app of different ML algorithms to predict digital and traditional medication adherence and diverse intervention outcomes. Positive results of these studies support and validate the feasibility of applying ML methods to predict user engagement in digital health apps such as a dDPP to improve patient adherence to digital therapeutics and, consequently, health outcomes. In concordance with the literature, we applied the most suitable algorithm for our data set (gradient-boosted forest), yielded highly accurate results for predicting digital adherence, and identified variables with the strongest contribution to our outcome to understand digital behaviors [22-26]. This paper described 2 ML models developed using weekly and daily dDPP engagement data. First, using the weekly dDPP vendor data set, we developed a weekly ML model, which was validated using the collected data from this dDPP study. On the basis of past activity patterns, the model yielded high precision and recall and accurately predicted patient engagement for the next week. However, a model trained with weekly patient data can only predict weekly engagement, limiting our ability to gain detailed insight into a patient’s behavior. Because an ideal model should be robust to different dynamics in patients’ engagement data, we then developed a daily ML model using the daily dDPP vendor data set, which incorporates additional attributes, including the type of meals logged per day and calories. The daily model also yielded high precision and recall values. This finding supports using such models to anticipate behavior, focusing on identifying low engagement to intervene before attrition.

In addition to calculating precision and recall for our models, we calculated the Shapley values for both types of models (weekly and daily) to further analyze and identify which variables contribute the most to overall prediction. Results from the Shapley values revealed that short-term frequency of activity engagement was the most informative feature in the daily and weekly data analyses, meaning that users were more likely to form and stick to short-term behavioral patterns than long-term patterns in the dDPP. This finding is consistent with a previous study on predicting exercise and steps [27]. Because of user propensity to engage in short-term behaviors, we considered...
the daily model for individual activities best suited to develop engagement profiles. Using variables with high Shapley values from the daily model, we successfully created distinct digital engagement phenotypes of dDPP users. This allows for further research into developing infrastructure for tailored messaging to increase and maintain engagement with active users and intervene against attrition for inactive users. Specifically, identifying high engagement, minimal engagement, and attrition with early dDPP use lends itself to determining individuals facing barriers to dDPP engagement and improving dDPP implementation. Identifying strengths and weaknesses within behavior phenotypes through our profiling methods can also inform what specific behaviors (ie, low-engagement behaviors) need to be targeted in messaging for a user’s success in using the dDPP.

Contributions and Implications

By leveraging digital behavioral usage data, we showed that we can successfully create digital engagement phenotypes, allowing for the future tailoring of digital health interventions based on patient needs. The methods used can extend beyond the prevention of metabolic disease, as an ML model incorporating behavioral usage variables can characterize prevention, maintenance, and wellness in other domains such as mental health, treatment adherence, and addiction prevention.

Limitations

The weekly data sets posed limitations to maximizing patient engagement through integrating ML into PAMS. A model trained using weekly data is limited to predict weekly dDPP engagement (limited scope of dDPP engagement). The weekly ML model did not provide enough granularity to be robust to different dynamics of app engagement (eg, a sudden drop in engagement in 1 week due to vacation or a suddenly busy day where the user does not log information). The high sensitivity in a weekly engagement model to unexpected changes in usage could, therefore, negatively impact the type of messaging and timely motivation delivered to the patient. Consequently, we shifted the prediction cycle for engagement by moving from a model based on weekly behavior to one based on daily behavior.

Data showed that the short-term frequency of various activities was the most informative feature, but the results could mean that our model is vulnerable to short-term disruption of user behavioral patterns. Consequently, although the weekly data-based and daily data-based models were sufficient to prove the feasibility of using ML approaches for predicting patient engagement, further development is needed to refine these models and include extra patient information. Improvements include (1) understanding potential errors in the model and data sets (eg, data set size; using vendor data sets is an imperfect representation of other dDPP interventions) and (2) reviewing initial hypotheses about the data set and the choice of algorithms. To build the refined model, we would benefit from more detailed data. In this case, we would need to replan attributes and test other ML algorithms to perform further model improvements.

Future Directions

With feasibility established, the next steps include creating user engagement phenotypes linked to personalized messaging interventions using behavior-based approaches to best motivate users to engage with the dDPP. We will also need to engineer the forest model and profile analysis to evolve as users change their engagement throughout participating in the dDPP so that messaging remains personalized to meet the users’ needs. Ultimately, this study demonstrated the potential value of ML and digital phenotyping to enhance the ability of digital behavior change interventions to predict engagement and personalize the interventions to maximize clinical impact.

Acknowledgments

Noom, Inc provided data from their commercial digital diabetes prevention program (dDPP) users, which was used as a baseline to train our machine learning model and obtain preliminary results. We thank Nina Singh for her feedback on the manuscript. This work was supported by the National Institute of Diabetes and Digestive and Kidney Diseases (grant 1R18DK118545-01A1; Principal Investigator: DMM). RVNV is funded by HRSA Ruth L. Kirschstein National Research Service Award (ID T32HP22238).

Authors’ Contributions

DVR, JC, and RVNV made substantial contributions to the conception or design of the work, as well as the acquisition, analysis, or interpretation of data for the work. They also contributed to drafting the work or revising it critically for important intellectual content. JC and RVNV contributed to the development of machine learning models and data analysis. All authors agreed to be accountable for all aspects of the work in ensuring that questions related to the accuracy or integrity of any part of the work are appropriately investigated and resolved. All authors gave their final approval of the version to be published.

Conflicts of Interest

None declared.

References


Abbreviations

AI: artificial intelligence  
AUPRC: area under the precision-recall curve  
AUROC: area under the receiver operating characteristic curve  
BIC: Bayesian information criterion  
dDPP: digital diabetes prevention program  
DPP: diabetes prevention program  
HIPAA: Health Insurance Portability and Accountability Act  
IRB: institutional review board  
LASSO: Least Absolute Shrinkage and Selection Operator  
LPA: latent profile analysis  
ML: machine learning  
PAMS: personalized automatic messaging system
Machine Learning Methods Using Artificial Intelligence Deployed on Electronic Health Record Data for Identification and Referral of At-Risk Patients From Primary Care Physicians to Eye Care Specialists: Retrospective, Case-Controlled Study

Joshua A Young, MD; Chin-Wen Chang, PhD; Charles W Scales, PhD; Saurabh V Menon, BTech; Chantal E Holy, MS, PhD; Caroline Adrienne Blackie, MS, OD, PhD

1Department of Ophthalmology, New York University School of Medicine, New York, NY, United States
2Data Science, Johnson & Johnson MedTech, Raritan, NJ, United States
3Medical and Scientific Operations, Johnson & Johnson Medtech, Vision, Jacksonville, FL, United States
4Mu Sigma Business Solutions Private Limited, Bangalore, India
5Epidemiology and Real-World Data Sciences, Johnson & Johnson MedTech, New Brunswick, NJ, United States
6Medical and Scientific Operations, Johnson & Johnson MedTech, Vision, Jacksonville, FL, United States

Corresponding Author:
Caroline Adrienne Blackie, MS, OD, PhD
Medical and Scientific Operations
Johnson & Johnson MedTech, Vision
7500 Centurion Parkway
Jacksonville, FL, 32256
United States
Phone: 1 9044331000
Email: cblackie@its.jnj.com

Abstract

Background: Identification and referral of at-risk patients from primary care practitioners (PCPs) to eye care professionals remain a challenge. Approximately 1.9 million Americans suffer from vision loss as a result of undiagnosed or untreated ophthalmic conditions. In ophthalmology, artificial intelligence (AI) is used to predict glaucoma progression, recognize diabetic retinopathy (DR), and classify ocular tumors; however, AI has not yet been used to triage primary care patients for ophthalmology referral.

Objective: This study aimed to build and compare machine learning (ML) methods, applicable to electronic health records (EHRs) of PCPs, capable of triaging patients for referral to eye care specialists.

Methods: Accessing the Optum deidentified EHR data set, 743,039 patients with 5 leading vision conditions (age-related macular degeneration [AMD], visually significant cataract, DR, glaucoma, or ocular surface disease [OSD]) were exact-matched on age and gender to 743,039 controls without eye conditions. Between 142 and 182 non-ophthalmic parameters per patient were input into 5 ML methods: generalized linear model, L1-regularized logistic regression, random forest, Extreme Gradient Boosting (XGBoost), and J48 decision tree. Model performance was compared for each pathology to select the most predictive algorithm.

Results: XGBoost demonstrated the best performance, showing, respectively, a prediction accuracy and an AUC of 78.6% (95% CI 78.3%-78.9%) and 0.878 for visually significant cataract, 77.4% (95% CI 76.7%-78.1%) and 0.858 for exudative AMD, 79.2% (95% CI 78.8%-79.6%) and 0.879 for nonexudative AMD, 72.2% (95% CI 69.9%-74.5%) and 0.803 for OSD requiring medication, 70.8% (95% CI 70.5%-71.1%) and 0.785 for glaucoma, 85.0% (95% CI 84.2%-85.8%) and 0.924 for type 1 nonproliferative diabetic retinopathy (NPDR), 82.2% (95% CI 80.4%-84.0%) and 0.911 for type 1 proliferative diabetic retinopathy (PDR), 81.3% (95% CI 81.0%-81.6%) and 0.891 for type 2 NPDR, and 82.1% (95% CI 81.3%-82.9%) and 0.900 for type 2 PDR.

Conclusions: The 5 ML methods deployed were able to successfully identify patients with elevated odds ratios (ORs), thus capable of patient triage, for ocular pathology ranging from 2.4 (95% CI 2.4-2.5) for glaucoma to 5.7 (95% CI 5.0-6.4) for type 1 NPDR, with an average OR of 3.9. The application of these models could enable PCPs to better identify and triage patients at risk for eye diseases.
risk for treatable ophthalmic pathology. Early identification of patients with unrecognized sight-threatening conditions may lead to earlier treatment and a reduced economic burden. More importantly, such triage may improve patients’ lives.

(Keywords: decision support for health professionals; tools, programs and algorithms; electronic health record; primary care; artificial intelligence; AI; prediction accuracy; triaging; AI model; eye care; ophthalmic)

Introduction

In the United States alone, more than 93 million adults were at high risk for vision loss in 2017; however, only 56.9% visited an eye care professional annually, and only 59.8% received a dilated eye examination [1]. More than 4 million Americans suffer from uncorrectable vision impairment, and more than 1 million are blind; this number is predicted to more than double by 2050 to 9 million due to the increasing epidemics of diabetes and other chronic diseases and our rapidly aging US population [2]. The impact of poor eyesight is manifest in its potentiation of comorbidities, particularly in increasing the risk of disability in patients with cognitive impairment [3]. Early identification of patients with unrecognized sight-threatening conditions may lead to earlier treatment and a reduced economic burden. More importantly, such triage may improve patients’ lives.

The identification and referral of patients at risk of vision loss from primary care practitioners (PCPs) to eye care professionals remains a challenge [4]. A 2010 study identified a number of barriers, including a lack of access to ophthalmic screening within the setting of the PCP’s office [4]. Some regional efforts have been made to improve the efficiency of triage of patients at risk for glaucoma [5] and diabetic retinopathy (DR) [6]; however, existing initiatives triage patients on only a few demographic and comorbidity parameters, whereas many systemic associations have been identified for age-related macular degeneration (AMD), cataract, DR, glaucoma, and ocular surface disease (OSD) [7–16].

Artificial intelligence (AI) modeling techniques are becoming increasingly important in ophthalmology in particular and medicine in general [17–20]. In ophthalmology, AI is used to calculate intraocular lens (IOL) powers [21–23], predict glaucoma progression [24,25], recognize DR [26], and classify ocular tumors [27]. To the best of our knowledge, AI has not yet been used to triage primary care patients for ophthalmology referral. In this study, the development, validation, and testing of multiple predictive machine learning (ML) methods for 5 leading sight-threatening and treatable ocular pathologies (i.e., AMD, visually significant cataract, DR, glaucoma, and OSD) that have the potential to be used by PCPs to triage patients, based on existing data in their electronic health records (EHRs), for referral to eye care specialists were reported.

Methods

AI Modeling

All AI techniques have in common the process of “training,” the adjustment of importance (i.e., weights) of attributes or intermediate values, based on a set of data referred to as a training set. The model performance is then assessed against another set of data called the test set. Similar model performance on training and test sets demonstrates model generalizability. The advent of large clinical databases has made possible the construction and training of both ML and neural network AI models. To this end, a large commercial EHR database that includes demographic, diagnostic, and therapeutic data to create and curate an ophthalmologically focused data set from which predictive models of multiple eye diseases can be built was used. We chose to compare several different ML methods to create models that might be used by PCPs to triage patients for referral to an eye care specialist. The models thus created used non-ophthalmic clinical and demographic data to assess relative risk scores for AMD, cataract, DR, glaucoma, and OSD.

Data Source

This retrospective, case-controlled study used data from the Optum deidentified EHR data set. EHRs provide efficient access to detailed patient-level longitudinal data that represent integral components of clinical care that may not necessarily be available through other retrospective database sources, such as administrative claims databases or patient registries [28,29]. The Optum EHR data set consists of data primarily from the United States and represents the clinical information of more than 80 million patients, including at least 7 million patients in each US census region from May 2000 to December 2019. Data from multiple EHR platforms, including Cerner, Epic, GE, and McKesson, are analyzed by Optum by means of natural language processing (NLP) to extract information about patient demographics, enrollment, diagnoses, biometrics, laboratory results, procedures, and medications [30]. The data set draws upon a network of more than 140,000 providers at more than 700 hospitals and 7000 clinics.

Ethical Considerations

The use of the Optum EHR data set was reviewed by the New England Institutional Review Board (IRB) and was determined to be exempt from broad IRB approval as this research project did not involve human subject research.

Outcome Measures

This study sought to predict the diagnosis of 5 major eye pathologies: AMD, cataract, DR, glaucoma, and OSD. The classification of AMD was based on the International Classification of Diseases, 10th Revision (ICD-10) codes and subdivided into nonexudative (H35.31%) and exudative (H35.32%) groups, in which “%” represents a wildcard. The classification of cataract required a more restrictive definition than simply H25%. Since no ICD-10 code distinguishes visually significant cataracts from those of lesser impact, we chose to use cataract surgery as a surrogate for visually significant...
cataract. For this study, cataract was defined by the cataract surgery Current Procedural Terminology (CPT) codes of 66982 and 66984 rather than by ICD-10. The classification of DR was based on the data set ICD-10 codes and subdivided into type 1 nonproliferative diabetic retinopathy (NPDR; H10.31%-H10.34%), type 1 proliferative diabetic retinopathy (PDR; H10.35%), type 2 NPDR (H11.31%-H11.34%), and type 2 PDR (H11.35%). Glaucoma was defined by the presence of 1 or more of 3 criteria: an ICD-10 code of H40.1% (open-angle glaucoma), the prescription of glaucoma medication, or the presence of a CPT code indicating glaucoma surgery. This definition was developed to capture not only patients with a recorded diagnosis of glaucoma but also those patients being treated for glaucoma or high-risk ocular hypertension for whom the diagnosis of glaucoma was not recorded in the data set. Similar to cataract, OSD required narrower criteria than simply H04.1% and H02.88% since these codes do not distinguish OSD requiring treatment from more mild presentations. For this study, OSD was defined rather restrictively as patients receiving cyclosporine opthalmic emulsion 0.05%, cyclosporine opthalmic solution 0.09%, or lifitegrast opthalmic solution 5% (see Tables 1 and 2).

**Table 1.** Listed medications for glaucoma.

<table>
<thead>
<tr>
<th>Type of medication</th>
<th>Examples</th>
</tr>
</thead>
<tbody>
<tr>
<td>Beta blockers</td>
<td>Levobunolol (Betagan, Akbeta), timolol (Timoptic, Betimal, Istalol), carteolol (Ocupress), metipranolol (Optipranolol), timolol gel (Timoptic Xe), betaxolol (Betoptic, Betoptic S)</td>
</tr>
<tr>
<td>Alpha agonists</td>
<td>Apraclonidine (Iopidine), brimonidine (Alphagan, Alphagan P), dipivefrin (Propine)</td>
</tr>
<tr>
<td>Carbonic anhydrase inhibitors</td>
<td>Dorzolamide (Trusopt), brinzolamide (Azopt)</td>
</tr>
<tr>
<td>Prostaglandin analogs</td>
<td>Latanoprost (Xalatan), bimatoprost 0.01% (% Lumigan), travoprost (Travatan Z), tafluprost (Zioptan), latanoprostene bunod (Vyzulta)</td>
</tr>
<tr>
<td>Prostaglandin analogs (combined medications)</td>
<td>Dorzolamide/timolol (Cosopt and Cospot P), brimonidine/timolol (Combigan), brinzolamide/brimonidine (Simbrinza), netarsudil/latanoprost (Rocklatan)</td>
</tr>
<tr>
<td>Rho kinase inhibitors</td>
<td>Netarsudil (Rhopressa)</td>
</tr>
</tbody>
</table>

**Table 2.** Listed procedures for glaucoma.

<table>
<thead>
<tr>
<th>ICD-10 code</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>0191T</td>
<td>Insertion of anterior segment aqueous drainage device, without extraocular reservoir, internal approach, into the trabecular meshwork; initial insertion</td>
</tr>
<tr>
<td>0253T</td>
<td>Insertion of anterior segment aqueous drainage device, without extraocular reservoir, internal approach, into the suprachoroidal space</td>
</tr>
<tr>
<td>0376T</td>
<td>Insertion of anterior segment aqueous drainage device, without extraocular reservoir, internal approach, into the trabecular meshwork; each additional device insertion (list separately in addition to code for primary procedure)</td>
</tr>
<tr>
<td>0449T</td>
<td>Insertion of aqueous drainage device, without extraocular reservoir, internal approach, into the subconjunctival space; initial device</td>
</tr>
<tr>
<td>0450T</td>
<td>Insertion of aqueous drainage device, without extraocular reservoir, internal approach, into the subconjunctival space; each additional device (list separately in addition to code for primary procedure)</td>
</tr>
<tr>
<td>0474T</td>
<td>Insertion of anterior segment aqueous drainage device, with creation of intraocular reservoir, internal approach, into the suprachoroidal space</td>
</tr>
<tr>
<td>65820</td>
<td>Goniotomy</td>
</tr>
<tr>
<td>65855</td>
<td>Trabeculectomy</td>
</tr>
<tr>
<td>66174</td>
<td>Transluminal dilation of aqueous outflow canal; without retention of device or stent</td>
</tr>
<tr>
<td>66175</td>
<td>Transluminal dilation of aqueous outflow canal; with retention of device or stent</td>
</tr>
<tr>
<td>66179</td>
<td>Aqueous shunt to extraocular equatorial plate reservoir, external approach; without graft</td>
</tr>
<tr>
<td>66180</td>
<td>Aqueous shunt to extraocular equatorial plate reservoir, external approach; with graft</td>
</tr>
<tr>
<td>66183</td>
<td>Insertion of anterior segment aqueous drainage device, without extraocular reservoir, external approach</td>
</tr>
<tr>
<td>66184</td>
<td>Revision of aqueous shunt to extraocular equatorial plate reservoir; without graft</td>
</tr>
<tr>
<td>66185</td>
<td>Revision of aqueous shunt to extraocular equatorial plate reservoir; with graft</td>
</tr>
<tr>
<td>66710</td>
<td>Ciliary body destruction by cyclophotocoagulation, trans-scleral approach</td>
</tr>
<tr>
<td>66711</td>
<td>Ciliary body destruction by cyclophotocoagulation, endoscopic approach (endoscopic cyclophotocoagulation)</td>
</tr>
</tbody>
</table>

ICD-10: International Classification of Diseases, 10th Revision.
Creation of Patient Cohorts

Five distinct cohorts (ocular cohorts) of patients (AMD n=294,739, cataract n=1,191,492, DR n=548,056, glaucoma n=843,560, and OSD n=660,218) were selected from the Optum EHR data set based on the aforementioned code definitions from October 2015 onward (to limit the analysis to the start of the ICD-10 coding system in the United States). The inclusion criteria were as follows: patients with diagnosis codes such as H3530% / H3531% / H3532%, H25%, E083%/E093%/E103%/E113%/E133%, H40%, or H041%/H0288% and EHRs with an ICD-10 diagnosis code type. Patients were excluded if they had an unknown birth year, were younger than 15 years, had less than 60 days of continuous enrollment in the database prior to their diagnosis, had a gender labeled as unknown, or had undergone a cataract-related procedure or diagnosis at baseline or not undergone a cataract-related procedure and diagnosis in the follow-up. Patients with multiple conditions (eg, glaucoma and OSD) were identified in both the glaucoma and OSD cohorts. For each patient, demographic information, complete clinical and drug use information, and comorbidities were identified. Multimedia Appendix 1 presents the patient inclusion and exclusion criteria and attrition data. All patients with the diagnoses present in the database during the specified inclusion period were considered for inclusion. Finally, the patients were segregated into subsets based on the AMD subtype or the DR subtype. In addition, only those patients who had open-angle glaucoma, had consumed a glaucoma-related medication, had undergone a glaucoma-related procedure in the follow-up, or had consumed dry eye and meibomian gland dysfunction (DEMGD)–related medications in the follow-up were retained. The final cohorts were as follows: exudative AMD n=32,072 (10.9%), nonexudative AMD n=114,839 (39%), cataract n=197,570 (66.6%), type I NPDR n=20,654 (5.9%), type I PDR n=4465 (1.3%), type II NPDR n=155,927 (44.8%), type II PDR n=21,032 (6%), glaucoma n=192,727 (22.8%), and OSD n=3720 (0.6%).

For each of the 5 cohorts, a control population was created from the pool of patients without ocular conditions. The control populations were matched 1:1 to each ocular cohort using exact matching on age and gender. A total of 743,039 patients with AMD, visually significant cataract, DR, glaucoma, or OSD were available in the Optum deidentified EHR data set, so these were exact-matched on age and gender to 743,039 controls without eye conditions.

Machine Learning

Several distinct ML approaches were followed to model the outcomes described earlier. These included the generalized linear model (GLM) [31], L1-regularized logistic regression (L1-LR) [32], random forest (RF) [33], Extreme Gradient Boosting (XGBoost) [34], and J48 decision tree (DT) [35].

Data Preprocessing

The data set consisted of 380 attributes, including demographic information, diagnoses, biometrics, laboratory results, procedures, and medications. Since some of these attributes, particularly some of the laboratory tests, were only sparsely represented, the data were pruned to remove attributes (ie, “features” in ML) with more than 20% missing values. Missing values were imputed with medians for continuous variables (eg, BMI), with a “Missing” group for categorical variables (eg, smoke or alcohol usage), and with the most frequent value for binary variables (eg, levels of lab test results). Winsorization of the data was performed to remove outliers and replace these with 0.1 and 99.9 percentile values. Further feature engineering was performed to remove or combine highly correlated features, such as “rheumatoid arthritis/collagen vascular disease” and its highly correlated cognate “connective tissue disease.” These feature engineering steps were performed individually for each case-controlled data set of each subpathology. The resultant data sets exhibited between 142 and 182 features after the above-described culling. The feature exclusion data sets for each of the 9 subpathologies were modeled using each of 5 distinct modeling strategies to produce a total of 45 individual ML models. These 45 models were produced and compared in a competitive fashion to identify the single-best model for each pathology.

Model Strategies

Logistic regression without regularization (LR), L1-LR, RF, and XGBoost models were performed in Python (3.8.5) using the Scikit-learn (0.23.2) and XGBoost (1.2.0) libraries. Next, 80% of the data were used for training, and 20% of the data were used for testing with 5-fold cross-validation. A grid search was used to optimize hyperparameters. For L1-LR, the regularization strength C was tuned. In the RF algorithm, the space of the number of trees and the maximum depth of each tree combination were searched. The hyperparameter tuning for XGBoost included the learning rate and the maximum depth of each tree. The ML modeling pipeline was established, and information of missing values fit and learned from the training data was applied to the test data set to avoid information leakage. J48 DT modeling, a Java-based implementation of the C4 tree, was performed in the WEKA ML workbench (University of Waikato). Finally, 10-fold cross-validation was used with an initial leaf size of 2% of the data set. The area under the curve (AUC) was assessed for all algorithms for each outcome to measure the overall performance of the binary classification models.

Results

Cohort Details

The demographic information of each cohort is shown in Table 3. Briefly, the total populations for modeling, for each cohort, varied in size from 7440 to 395,140. Populations were mostly female for AMD, cataract, glaucoma, and OSD requiring medications, and the average age ranged from 51 to 80 years.

The performance of different ML strategies varied as well (Figures 1 and 2 and Table 4), but in all cases, XGBoost demonstrated the best performance, showing, respectively, a prediction accuracy and an AUC of 78.6% (95% CI 78.3%-78.9%) and 0.878 for visually significant cataract, 77.4% (95% CI 76.7%-78.1%) and 0.858 for exudative AMD, 79.2% (95% CI 78.8%-79.6%) and 0.879 for nonexudative AMD, 72.2% (95% CI 69.9%-74.5%) and 0.803 for OSD requiring medication, 70.8% (95% CI 70.5%-71.1%) and 0.785 for "..."
glaucoma, 85.0% (95% CI 84.2%-85.8%) and 0.924 for type 1 NPDR, 82.2% (95% CI 80.4%-84.0%) and 0.911 for type 1 PDR, 81.3% (95% CI 81.0%-81.6%) and 0.891 for type 2 NPDR, and 82.1% (95% CI 81.3%-82.9%) and 0.900 for type 2 PDR (Table 4). XGBoost identified several clinical attributes that were important for diagnosis prediction (Figure 3).

The top-performing models identified the following clinical and demographic features that were primarily contributing to the predictions for each pathology (Figure 3; continuous measures showed positive associations):

- Exudative AMD diagnosis prediction was associated, in order of importance, with average household income, percentage college education, geographical division (Middle Atlantic, East North Central, East South Central, New England, South Atlantic/West South Central, Mountain, West North Central, Pacific, other/unknown), the BMI, and the Elixhauser score (comorbidity index).
- Nonexudative AMD demonstrated similar associations. In order of importance, these were average household income, percentage college education, region (Northeast, Midwest, South, West, other/unknown), smoking, and the Elixhauser score.
- Glaucoma clinical associations, in order of importance, included average household income, percentage college education, adrenal or androgen use, the BMI, and race.
- Cataract clinical associations, in order of importance, included average household income, percentage college education, region, the BMI, and smoking.
- OSD associations, in order of importance, included average household income, percentage college education, geographical division, rheumatoid arthritis and connective tissue disease, and region.
- DR associations varied over different subpathologies but generally included the Elixhauser score, high serum glucose, the BMI, hypertension, chronic pulmonary disease, depression, cardiac arrhythmia, and obesity.

Performance in predicting the presence of pathology ranged from 71% in the case of glaucoma to 87% in the case of type I PDR, with an average performance of 80% across all groups. Since the intent was to identify at-risk patients, these performance values were used to determine disease odds ratios (ORs) according to the method described by Hogue et al [36]. Applying this to each of the models provided a clinically useful measure. The models identified patients with elevated ORs of the prevalence of pathology from 2.4 in the case of glaucoma to 5.7 in the case of type I NPDR, with an average OR of 3.9 (Table 5).
Table 3. Demographic information of each cohort with ocular disease. For each cohort, a control (age- and gender-matched) population of similar size was generated, without the condition of interest.

<table>
<thead>
<tr>
<th>Characteristic</th>
<th>Exudative AMD(^a) (n=32,072)</th>
<th>Nonexudative AMD (n=148,897)</th>
<th>Cataract (n=197,570)</th>
<th>OSD(^b) requiring medication (n=3720)</th>
<th>Glaucoma (n=192,727)</th>
<th>Type I NPDR(^c) (n=20,654)</th>
<th>Type I PDR(^d) (n=4465)</th>
<th>Type II NPDR (n=155,927)</th>
<th>Type II PDR (n=21,032)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Age (years), mean (SD)</td>
<td>79.8 (10.4)</td>
<td>77.1 (10.7)</td>
<td>69.7 (9.9)</td>
<td>68.3 (14.0)</td>
<td>72.4 (13.3)</td>
<td>51.5 (16.0)</td>
<td>52.1 (14.6)</td>
<td>64.4 (12.9)</td>
<td>61.6 (12.7)</td>
</tr>
<tr>
<td>Gender (female), n (%)</td>
<td>19,885 (62.0)</td>
<td>70,971 (61.8)</td>
<td>115,183 (58.3)</td>
<td>3050 (82.0)</td>
<td>108,698 (56.4)</td>
<td>10,203 (49.4)</td>
<td>2170 (48.6)</td>
<td>77,028 (49.4)</td>
<td>10,032 (47.7)</td>
</tr>
<tr>
<td>Race, n (%)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Asian</td>
<td>353 (1.1)</td>
<td>1608 (1.4)</td>
<td>3951 (2.0)</td>
<td>52 (1.4)</td>
<td>3662 (1.9)</td>
<td>186 (0.9)</td>
<td>31 (0.7)</td>
<td>4054 (2.6)</td>
<td>484 (2.3)</td>
</tr>
<tr>
<td>Black</td>
<td>374 (2.1)</td>
<td>2756 (2.4)</td>
<td>13,632 (6.9)</td>
<td>272 (7.3)</td>
<td>30,065 (15.6)</td>
<td>2231 (10.8)</td>
<td>545 (12.2)</td>
<td>24,948 (16.0)</td>
<td>3912 (18.6)</td>
</tr>
<tr>
<td>White</td>
<td>27,903 (87.0)</td>
<td>97,843 (85.2)</td>
<td>160,229 (81.1)</td>
<td>3281 (88.2)</td>
<td>139,342 (72.3)</td>
<td>16,337 (79.1)</td>
<td>3393 (76.0)</td>
<td>106,342 (68.2)</td>
<td>13,166 (62.6)</td>
</tr>
<tr>
<td>Unknown</td>
<td>3143 (9.8)</td>
<td>12,632 (11.0)</td>
<td>23,511 (11.9)</td>
<td>112 (3.0)</td>
<td>19,658 (10.2)</td>
<td>1900 (9.2)</td>
<td>500 (11.2)</td>
<td>20,582 (13.2)</td>
<td>3449 (16.4)</td>
</tr>
<tr>
<td>Ethnicity, n (%)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Hispanic</td>
<td>513 (1.6)</td>
<td>2067 (1.8)</td>
<td>5927 (3.0)</td>
<td>86 (2.3)</td>
<td>7516 (3.9)</td>
<td>888 (4.3)</td>
<td>223 (5.0)</td>
<td>13,722 (8.8)</td>
<td>2608 (12.4)</td>
</tr>
<tr>
<td>Non-Hispanic</td>
<td>27,774 (86.6)</td>
<td>96,465 (84.0)</td>
<td>168,132 (85.1)</td>
<td>3553 (95.5)</td>
<td>164,589 (85.4)</td>
<td>17,804 (86.2)</td>
<td>3764 (84.3)</td>
<td>124,118 (79.6)</td>
<td>15,900 (75.6)</td>
</tr>
<tr>
<td>Unknown</td>
<td>3784 (11.8)</td>
<td>16,307 (14.2)</td>
<td>23,511 (11.9)</td>
<td>82 (2.2)</td>
<td>20,622 (10.7)</td>
<td>1962 (9.5)</td>
<td>478 (10.7)</td>
<td>18,088 (11.6)</td>
<td>2524 (12.0)</td>
</tr>
<tr>
<td>Education (college educated), n (%)</td>
<td>7761 (24.2)</td>
<td>27,906 (24.3)</td>
<td>47,614 (24.1)</td>
<td>868 (23.2)</td>
<td>47,411 (24.6)</td>
<td>4936 (23.9)</td>
<td>1058 (23.7)</td>
<td>37,111 (23.8)</td>
<td>4943 (23.5)</td>
</tr>
<tr>
<td>Size of control population, n</td>
<td>32,072</td>
<td>114,839</td>
<td>197,570</td>
<td>3720</td>
<td>192,727</td>
<td>20,654</td>
<td>4465</td>
<td>155,927</td>
<td>21,032</td>
</tr>
<tr>
<td>Total population for modeling (cohort+control), n</td>
<td>64,144</td>
<td>229,678</td>
<td>395,140</td>
<td>7440</td>
<td>385,454</td>
<td>41,308</td>
<td>8930</td>
<td>311,854</td>
<td>42,064</td>
</tr>
</tbody>
</table>

\(^a\)AMD: age-related macular degeneration.  
\(^b\)OSD: ocular surface disease.  
\(^c\)NPDR: nonproliferative diabetic retinopathy.  
\(^d\)PDR: proliferative diabetic retinopathy.
Figure 1. Model accuracy by pathology degeneration; AUC = area under the curve; CI = confidence interval; J48 = Decision tree; LR = Logistic Regression without regularization; LR-L1 = L1-regularized logistic regression; NPDR = non-proliferative diabetic retinopathy; OSD = ocular surface disease; PDR = proliferative diabetic retinopathy; XGB = XGBoost.

Figure 2. Receiver operating characteristic (ROC) curves illustrating the diagnostic ability of the models for the 9 pathologies. amd: age-related macular degeneration; auc: area under the curve; demgd: dry eye and meibomian gland dysfunction; j48: decision tree; l1: L1-regularized logistic regression; lr: logistic regression without regularization; npdr: nonproliferative diabetic retinopathy; pdr: proliferative diabetic retinopathy; rf: random forest; xgb: Extreme Gradient Boosting.
Table 4. Model accuracy, AUC, sensitivity, and specificity.

<table>
<thead>
<tr>
<th>Outcome and algorithms</th>
<th>Accuracy (95% CI)</th>
<th>AUC (95% CI)</th>
<th>Sensitivity</th>
<th>Specificity</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Cataract</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>XGBoost</td>
<td>78.6% (78.3%-78.9%)</td>
<td>0.878 (0.875-0.880)</td>
<td>0.796</td>
<td>0.776</td>
</tr>
<tr>
<td>RF</td>
<td>72.1% (71.8%-72.4%)</td>
<td>0.811 (0.808-0.814)</td>
<td>0.749</td>
<td>0.693</td>
</tr>
<tr>
<td>LR-L1</td>
<td>68.9% (68.6%-69.2%)</td>
<td>0.767 (0.764-0.771)</td>
<td>0.683</td>
<td>0.695</td>
</tr>
<tr>
<td>LR</td>
<td>68.9% (68.6%-69.2%)</td>
<td>0.767 (0.764-0.771)</td>
<td>0.683</td>
<td>0.695</td>
</tr>
<tr>
<td>J48 DT</td>
<td>66.5% (N/A)</td>
<td>0.710 (N/A)</td>
<td>0.702</td>
<td>0.628</td>
</tr>
<tr>
<td><strong>Exudative AMD</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>XGBoost</td>
<td>77.4% (76.7%-78.1%)</td>
<td>0.858 (0.851-0.863)</td>
<td>0.769</td>
<td>0.778</td>
</tr>
<tr>
<td>RF</td>
<td>73.0% (72.2%-73.8%)</td>
<td>0.817 (0.810-0.825)</td>
<td>0.745</td>
<td>0.715</td>
</tr>
<tr>
<td>LR-L1</td>
<td>71.8% (71.0%-72.6%)</td>
<td>0.794 (0.786-0.802)</td>
<td>0.716</td>
<td>0.720</td>
</tr>
<tr>
<td>LR</td>
<td>71.8% (71.0%-72.6%)</td>
<td>0.794 (0.786-0.801)</td>
<td>0.717</td>
<td>0.720</td>
</tr>
<tr>
<td>J48 DT</td>
<td>68.1% (N/A)</td>
<td>0.721 (N/A)</td>
<td>0.707</td>
<td>0.660</td>
</tr>
<tr>
<td><strong>Nonexudative AMD</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>XGBoost</td>
<td>79.2% (78.8%-79.6%)</td>
<td>0.879 (0.876-0.882)</td>
<td>0.801</td>
<td>0.783</td>
</tr>
<tr>
<td>RF</td>
<td>73.3% (72.9%-73.7%)</td>
<td>0.823 (0.820-0.827)</td>
<td>0.768</td>
<td>0.698</td>
</tr>
<tr>
<td>LR-L1</td>
<td>71.3% (70.9%-71.7%)</td>
<td>0.794 (0.790-0.798)</td>
<td>0.729</td>
<td>0.697</td>
</tr>
<tr>
<td>LR</td>
<td>71.3% (70.9%-71.7%)</td>
<td>0.794 (0.790-0.798)</td>
<td>0.727</td>
<td>0.700</td>
</tr>
<tr>
<td>J48 DT</td>
<td>68.1% (N/A)</td>
<td>0.725 (N/A)</td>
<td>0.741</td>
<td>0.622</td>
</tr>
<tr>
<td><strong>OSD</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>XGBoost</td>
<td>72.2% (69.9%-74.5%)</td>
<td>0.803 (0.780-0.824)</td>
<td>0.708</td>
<td>0.735</td>
</tr>
<tr>
<td>RF</td>
<td>70.9% (68.6%-73.2%)</td>
<td>0.771 (0.747-0.795)</td>
<td>0.749</td>
<td>0.669</td>
</tr>
<tr>
<td>LR-L1</td>
<td>69.0% (66.7%-71.3%)</td>
<td>0.757 (0.732-0.782)</td>
<td>0.691</td>
<td>0.688</td>
</tr>
<tr>
<td>LR</td>
<td>69.5% (67.2%-71.8%)</td>
<td>0.757 (0.733-0.782)</td>
<td>0.688</td>
<td>0.702</td>
</tr>
<tr>
<td>J48 DT</td>
<td>65.1% (N/A)</td>
<td>0.702 (N/A)</td>
<td>0.675</td>
<td>0.628</td>
</tr>
<tr>
<td><strong>Glaucoma</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>XGBoost</td>
<td>70.8% (70.5%-71.1%)</td>
<td>0.785 (0.782-0.788)</td>
<td>0.689</td>
<td>0.728</td>
</tr>
<tr>
<td>RF</td>
<td>67.9% (67.6%-68.2%)</td>
<td>0.741 (0.738-0.745)</td>
<td>0.656</td>
<td>0.702</td>
</tr>
<tr>
<td>LR-L1</td>
<td>61.8% (61.5%-62.1%)</td>
<td>0.669 (0.665-0.673)</td>
<td>0.622</td>
<td>0.614</td>
</tr>
<tr>
<td>LR</td>
<td>61.8% (61.5%-62.1%)</td>
<td>0.669 (0.665-0.673)</td>
<td>0.619</td>
<td>0.617</td>
</tr>
<tr>
<td>J48 DT</td>
<td>62.0% (N/A)</td>
<td>0.647 (N/A)</td>
<td>0.647</td>
<td>0.593</td>
</tr>
<tr>
<td><strong>Type I NPDR</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>XGBoost</td>
<td>85.0% (84.2%-85.8%)</td>
<td>0.924 (0.919-0.930)</td>
<td>0.850</td>
<td>0.850</td>
</tr>
<tr>
<td>RF</td>
<td>79.5% (78.6%-80.4%)</td>
<td>0.872 (0.864-0.879)</td>
<td>0.799</td>
<td>0.790</td>
</tr>
<tr>
<td>LR-L1</td>
<td>83.5% (82.7%-84.3%)</td>
<td>0.908 (0.902-0.915)</td>
<td>0.847</td>
<td>0.824</td>
</tr>
<tr>
<td>LR</td>
<td>83.5% (82.7%-84.3%)</td>
<td>0.908 (0.902-0.915)</td>
<td>0.847</td>
<td>0.824</td>
</tr>
<tr>
<td>J48 DT</td>
<td>73.8% (N/A)</td>
<td>0.796 (N/A)</td>
<td>0.756</td>
<td>0.721</td>
</tr>
<tr>
<td><strong>Type I PDR</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>XGBoost</td>
<td>82.2% (80.4%-84.0%)</td>
<td>0.911 (0.897-0.924)</td>
<td>0.816</td>
<td>0.828</td>
</tr>
<tr>
<td>RF</td>
<td>77.3% (75.4%-79.2%)</td>
<td>0.861 (0.846-0.878)</td>
<td>0.802</td>
<td>0.744</td>
</tr>
<tr>
<td>Outcome and algorithms</td>
<td>Accuracy (95% CI)</td>
<td>AUC (95% CI)</td>
<td>Sensitivity</td>
<td>Specificity</td>
</tr>
<tr>
<td>------------------------</td>
<td>-------------------</td>
<td>--------------</td>
<td>-------------</td>
<td>-------------</td>
</tr>
<tr>
<td>LR-L1</td>
<td>81.2% (79.4%-83.0%)</td>
<td>0.895 (0.881-0.910)</td>
<td>0.847</td>
<td>0.777</td>
</tr>
<tr>
<td>LR</td>
<td>80.8% (79.0%-82.6%)</td>
<td>0.894 (0.880-0.910)</td>
<td>0.829</td>
<td>0.787</td>
</tr>
<tr>
<td>J48 DT</td>
<td>72.4% (N/A)</td>
<td>0.804 (N/A)</td>
<td>0.761</td>
<td>0.686</td>
</tr>
</tbody>
</table>

**Type II NPDR**

<table>
<thead>
<tr>
<th>Algorithms</th>
<th>AUC (95% CI)</th>
<th>Sensitivity</th>
<th>Specificity</th>
</tr>
</thead>
<tbody>
<tr>
<td>XGBoost</td>
<td>81.3% (81.0%-81.6%)</td>
<td>0.845</td>
<td>0.782</td>
</tr>
<tr>
<td>RF</td>
<td>75.1% (74.8%-75.4%)</td>
<td>0.751</td>
<td>0.752</td>
</tr>
<tr>
<td>LR-L1</td>
<td>79.1% (78.8%-79.4%)</td>
<td>0.843</td>
<td>0.739</td>
</tr>
<tr>
<td>LR</td>
<td>79.1% (78.8%-79.4%)</td>
<td>0.844</td>
<td>0.739</td>
</tr>
<tr>
<td>J48 DT</td>
<td>69.6% (N/A)</td>
<td>0.635</td>
<td>0.757</td>
</tr>
</tbody>
</table>

**Type II PDR**

<table>
<thead>
<tr>
<th>Algorithms</th>
<th>AUC (95% CI)</th>
<th>Sensitivity</th>
<th>Specificity</th>
</tr>
</thead>
<tbody>
<tr>
<td>XGBoost</td>
<td>82.1% (81.3%-82.9%)</td>
<td>0.841</td>
<td>0.801</td>
</tr>
<tr>
<td>RF</td>
<td>77.7% (76.8%-78.6%)</td>
<td>0.763</td>
<td>0.790</td>
</tr>
<tr>
<td>LR-L1</td>
<td>79.9% (79.0%-80.8%)</td>
<td>0.834</td>
<td>0.763</td>
</tr>
<tr>
<td>LR</td>
<td>80.0% (79.1%-80.9%)</td>
<td>0.847</td>
<td>0.753</td>
</tr>
<tr>
<td>J48 DT</td>
<td>71.1% (N/A)</td>
<td>0.674</td>
<td>0.748</td>
</tr>
</tbody>
</table>

---

*a* AUC: area under the curve.  
*b* XGBoost: Extreme Gradient Boosting.  
*c* RF: random forest. 
*d* L1-LR: L1-regularized logistic regression.  
*e* LR: logistic regression without regularization.  
*f* DT: decision tree.  
*g* N/A: not applicable.  
*h* AMD: age-related macular degeneration.  
*i* OSD: ocular surface disease.  
*j* NPDR: nonproliferative diabetic retinopathy.  
*k* PDR: proliferative diabetic retinopathy. 

**Figure 3.** Clinical features primarily contributing to the predictions for each pathology. amd: age-related macular degeneration; demgd: dry eye and meibomian gland dysfunction; hh: household; npdr: nonproliferative diabetic retinopathy; pct: percentage; pdr: proliferative diabetic retinopathy; xgb: Extreme Gradient Boosting.
Table 5. Model accuracy and ORs by pathology.

<table>
<thead>
<tr>
<th>Pathology</th>
<th>Model accuracy, %</th>
<th>OR (95% CI)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Exudative AMD</td>
<td>77</td>
<td>3.4 (3.2-3.7)</td>
</tr>
<tr>
<td>Nonexudative AMD</td>
<td>79</td>
<td>3.8 (3.6-4.0)</td>
</tr>
<tr>
<td>Cataract</td>
<td>79</td>
<td>3.7 (3.6-3.8)</td>
</tr>
<tr>
<td>OSD</td>
<td>72</td>
<td>2.6 (2.1-3.3)</td>
</tr>
<tr>
<td>Glaucoma</td>
<td>71</td>
<td>2.4 (2.4-2.5)</td>
</tr>
<tr>
<td>Type I PDR</td>
<td>82</td>
<td>4.6 (3.6-5.9)</td>
</tr>
<tr>
<td>Type I NPDR</td>
<td>85</td>
<td>5.7 (5.0-6.4)</td>
</tr>
<tr>
<td>Type II PDR</td>
<td>82</td>
<td>4.6 (4.1-5.1)</td>
</tr>
<tr>
<td>Type II NPDR</td>
<td>81</td>
<td>4.3 (4.2-4.5)</td>
</tr>
</tbody>
</table>

aOR: odds ratio.
bAMD: age-related macular degeneration.
cOSD: ocular surface disease.
dPDR: proliferative diabetic retinopathy.
eNPDR: nonproliferative diabetic retinopathy.

Discussion

Principal Findings

A major challenge of current deep learning (DL) models is that their training requires a large amount of data because insufficient data may decrease the performance of DL models [37]. The original EHR data pool for this study comprised more than 80 million patients, one of the largest AI projects of its kind in ophthalmology. The final study populations totaled 1,486,078 patients, 50% of whom were controls. In addition to the substantial patient population, this study examined 9 subpathologies using 5 different analytical modeling approaches to identify the most predictive model for each pathology.

The goal of this effort was to create a digital health tool to identify patients at higher risk for the presence of ophthalmic pathology and to do this based solely on the sort of non-ophthalmic data to which a PCP would have access. The authors do not propose to either make definitive ophthalmic diagnoses or predict the development of future pathology. Rather, this work seeks to identify patients whose clinical and demographic context is associated with the presence of AMD, cataract, clinically significant DR, glaucoma, or OSD of a magnitude requiring pharmacological therapy. The creation, demonstration, and real-world validation (within a clinical setting) of a deployable digital tool will be the next step of this project.

The application of such a model in the clinical setting would allow a PCP to identify patients nearly 4 times more likely to have ophthalmic pathology. Such a tool would bring a substantial benefit in the triage and referral of at-risk patients to eye care professionals.

Data and Outcome Engineering

These data consist of diagnostic and procedure codes; biometric data, such as the BMI and vital signs; demographic information, including socioeconomic and geographical information; laboratory results; and medications prescribed. This information does not include the physician notes that might provide a rationale for the diagnoses recorded. Indeed, since only a limited number of diagnoses may be listed on a claim, it is possible that some extant diagnoses may have gone unrecorded. However, diagnoses like cataract and OSD may be overrepresented since the ICD-10 taxonomy does not distinguish between clinically significant cataract and OSD from cases in which these pathologies are subclinical. Indeed, it would be of little clinical utility to build an AI model that detects subclinical cataracts.

Ours is not the first study to be faced with the challenge of identifying clinically relevant diagnoses from large data sets. A 2018 study [38] investigated the precision of ICD-10 codes for patients with uveitis and found that 13 of 27 uveitides were imprecisely defined and that multiple codes were used to describe the same pathology. A 2020 study of ocular pathology in patients with stroke [39] noted fewer patients with glaucoma than anticipated and attributed this to the lack of ophthalmology clinic data. The authors noted that patients may be on glaucoma medications without a concurrent ICD-10 code recorded for glaucoma, suggesting that a diagnosis of glaucoma may have been recorded in the patients’ medical records before incorporation into the data set. The authors sought, therefore, to define the glaucoma cohort as those patients who met 1 or more of 3 criteria: an ICD-10 code of H40.1% (open-angle glaucoma), the prescription of glaucoma medication, or the presence of a CPT code indicating glaucoma surgery (see Tables 1 and 2). This definition was developed to both detect glaucoma patients without glaucoma ICD-10 codes and to exclude patients inappropriately labeled as glaucoma by ICD-10. This definition resulted in a substantial winnowing of the glaucoma cohort from 1,368,700, 50% of whom were controls, to 385,514 patients.

The authors took a similar approach to the cataract and OSD study populations. Cataract and OSD are among the most frequently recorded diagnoses on claims [40]. Cataract, in
particular, is nearly ubiquitous in elderly patients and was the most common ophthalmic ICD-10 diagnosis of those examined here. Since only a subset of these patients require cataract surgery, the detection of cataract alone is not clinically useful. ICD-10 coding does not distinguish between cataracts requiring surgery and those that do not. However, CPT coding, in a sense does make this distinction. Therefore, CPT codes of 66984 (cataract extraction with intraocular lens) and 66982 (complex cataract extraction) were chosen as the criteria for clinically significant cataracts. This narrowing of the inclusion criteria reduced our cataract study population from 2,087,836, 50% of whom were controls, to 395,140 patients. OSD coding is even more problematic. A large number of ICD-10 codes are available, and clinical significance is difficult to establish. Our initial cohort of OSD patients and controls totaled 1,182,912 patients. To model the clinical context associated with OSD, a restrictive criterion was chosen: the prescription of topical cyclosporine or lifitegrast. This greatly reduced the OSD population to 7440 patients, but this ensured the final population represented patients with clinically meaningful disease. No outcome engineering measures were applied to the AMD groups or to the DR groups, each of which was defined by its corresponding ICD-10 code.

In addition, PDR and NPDR could have been combined into 1 group since the referring physician probably would not care about what sort of DR the patient has. However, the NPDR group is so much larger than the PDR group that the authors do not expect that the segmentation is detrimental.

Clinical and Demographic Attributes and Feature Engineering

The initial data set included a large number of attributes or “features” (in the language of ML), totaling 380 individual parameters. To produce models that would not be burdensome for the clinician to use, the authors sought to reduce the number of attributes required by each model. This reduction and modification of model parameters is referred to as “feature engineering.” For a feature to be included in the final model, several criteria needed to be met. The feature must play a significant role in the model’s outcome. It is self-evident that features that do not contribute substantially to a model may be discarded with little impact on model performance. In the case of the XGBoost models, parameter optimization was performed by the grid search algorithm [41]. The second feature inclusion criterion was noncorrelation with other features. In some cases, such as between weight and the BMI, the correlation is evident. However, the correlation between other clinical features only becomes clear on analysis. The issue of feature correlation highlights a difference between AI and traditional risk analysis studies. When studied individually, certain attributes, such as obesity and socioeconomic status, may be identified as disease risk factors. However, when viewed collectively, the importance of 1 of these may be reduced if the 2 attributes are highly correlated. The third criterion for feature inclusion was high frequency in the data set. Some of the laboratory values, particularly serum fibrinogen, were so sparse in the data set that exclusion of the feature was preferable to the alternatives of sample reduction or interpolation. Two thresholds for feature sparsity were established in this project. Models were built upon data sets that excluded features with more than 20% missing values. Feature engineering substantially benefits from guidance by clinical domain experts [42], and our feature and outcome engineering was clinically informed, particularly in the realm of the diagnostic criteria described earlier. The features included in the final XGBoost model, the top-performing strategy, are available as supplementary materials to this manuscript. XGBoost is a DT-based ensemble modeling method. It can effectively capture the nonlinear relationship between predictors and the outcome by combining many weaker models to create a strong model. “Weak” and “strong” here refer to how correlated the models are to the outcome. The algorithm added models sequentially, and the next model corrected the error from the previous model. Through this iterative process, the data can be eventually accurately predicted by the model.

Usage Data and Generalizability

The application of usage data to this effort is both a weakness and a strength of this project. These data do not contain the richness of a complete medical record. It is therefore impossible to establish the criteria under which the clinicians made the diagnoses recorded—hence our outcome engineering maneuvers to establish stricter criteria (eg, using CPT codes for cataract surgery to identify patients with clinically significant cataract). At the same time, models built upon these sorts of data are more generalizable and available than models built upon more specific and perhaps more idiosyncratic data sources. These are precisely the sorts of data available to PCPs, making these models more easily deployable than models built upon a specific medical record system. Indeed, the availability of these data is illustrated by our being able to investigate a base of more than 80 million patients from disparate health care systems.

Definitions of the parameters used in these models is a topic worth addressing. The parameters ingested by the models that are used to make predictions include pathologies and demographics that would ordinarily require a clear and consistent definition. These parameters include macular degeneration whose definition should be established a priori to demographic terms, such as gender and sex, that not only require definition but also incorporate the idea of nonbinary values.

It is the nature of large electronic medical record studies that such definitions are impossible to impose externally and that the interpretations of gender, hypertension, diabetes, and glaucoma are likely to vary among the practitioners and patients who themselves may be the source of the data of these values in the data set. Our use of a database of 80 million patients provides a large degree of protection from selection bias. However, because these clinical definitions are intrinsic to the data set itself, a great deal of caution must be exercised when attempting to draw inferences about pathogenesis simply by evaluating the most correlative features of the model. However, the limitation of the model to revealing the disease process makes the model no less valuable in its ability to predict which patients are at the highest risk for unrecognized eye disease.

Hierarchical Relationships

It should be noted that the clinical features identified as relevant by each of the pathology models should be viewed as correlative...
but not necessarily causative. It is better to think of the collection of clinical values as a patient’s clinical milieu rather than as a collection of individual risk factors. Although it is difficult to imagine that college education is itself a risk factor for pathology, its correlation and importance to a given model should not be discounted, since it does contribute to the model’s predictiveness of the presence of pathology. All of this is not to say that causation may not exist in the relation between some of these features and the pathologies modeled. Highly multidimensional clinical AI studies like this one may identify previously unrecognized factors that directly influence pathogenesis. However, causative connection cannot be established by this sort of study and would require a more traditional experimental approach. Although the J48 DT models did not perform as well as the GLM or XGBoost strategies, they are informative in that they describe hierarchical relationships among clinical features. As an example, the J48 model for glaucoma identifies race, systemic steroids, and antidiabetic medication use as important clinical features. However, the model dictates the order in which these factors should be considered, assessing race only after it is established whether the patient takes antidiabetic medications and assessing systemic steroid use only after these first 2 attributes have been determined. Such a hierarchical relationship among clinical features and demographic characteristics would be enormously difficult to establish in traditional reduced-dimensional scientific queries. This gestalt approach to multidimensional clinical context is one of the strengths of AI.

Decision Support

Ophthalmology is well suited for AI, given the rich visual information and data available; complex ophthalmological systems are better understood and eye care enhanced through sophisticated analysis and prediction. Integrating AI into clinical practice may facilitate better patient outcomes, given the complexity of disease diagnosis, treatment selection, and clinical testing. Ophthalmological clinical decision support systems that aid in diagnosis could improve the accuracy and efficiency of decision-making processes in ophthalmology, ultimately leading to improved patient access, outcomes, and potentially costs [43].

These models predict the presence of extant pathology. They would be of value in the identification of populations in which these pathologies are substantially more prevalent than in the general population. The models should not be used to make a diagnosis for an individual patient but rather to identify patients at risk of having undetected AMD, cataract, DR, glaucoma, or OSD. Further, these models are built upon clinical data in which an ophthalmic pathology is or is not present. That is to say, the models presented here are not constructed to predict the development of future pathology. It may or may not be the case that a particular clinical context, as defined by the multidimensional features incorporated into the models, may predict the development of future disease, but that is not appropriate way to use the models presented. These models predict the presence of ophthalmic pathology based upon non-ophthalmic data and would be best used for triage and referrals from non-ophthalmologists to eye care specialists. The research is designed to raise awareness about the variables associated with referral to heighten PCPs’ vigilance to the clinical and demographic characteristics that may need further reflection and attention.

Real-World Application Prospects of Ophthalmological AI Models

Advances in computing power combined with disruptions in health care resulting from unprecedented circumstances of the COVID-19 pandemic have prompted the worldwide exploration of AI-based systems in several medical subfields, including ophthalmology [44]. Ophthalmology has been at the forefront of AI research, in particular ML and DL approaches, because of the ubiquitous availability of noninvasive, rapid, and relatively inexpensive ophthalmic imaging [45]. Ophthalmic AI systems are advantageous in that they decrease the amount of time required to interpret image data, enable ophthalmologists to gain a greater understanding of disease progression, and assist with early-stage diagnosis, staging, and prognosis [46].

Numerous factors will determine the successful adoption of AI technologies into clinical practice. AI innovations that help clinicians manage the complexity (rather than add yet another layer of complexity) associated with effective ophthalmological care will likely be better received. In addition, the ability for critical appraisals by optometrists and ophthalmologists will be key to validating the theoretic models. AI models can be difficult to interpret and explain, which can make it difficult for stakeholders to understand how decisions are made [47]. It is important that the AI models be transparent and explainable in order to gain and maintain the trust of health care professionals, patients, and other decision makers. Providers of AI technologies and educators also need to ensure that training needs are adequately assessed and value to patient outcomes demonstrated if the promise of AI in ophthalmological care is to be realized.

AI has the potential to provide invaluable insights across multiple domains of ophthalmology. By leveraging ML algorithms, AI can process and analyze vast amounts of information, including physiological data, EHRs, 3D images, radiology images, histologic evaluation, genomic sequencing, and administrative and billing data. One advantage that could be realized by the algorithms discussed herein is that they use commonly collected data contained within an EHR system to identify eye disease risk. This means that the algorithms could be deployed in the background of an EHR to enable inference of an entire PCP’s or practice’s patient population. The results of this inference could appear as a flag in a patient chart, alerting the PCP for a given patient as to the need to refer to an eye care professional for further evaluation. The approach of deploying these algorithms within the EHR would also enable further validation and assessment of algorithm generalizability prior to clearing the algorithm for regular use by PCPs. Additional validation steps such as this would help identify any local biases for a given patient population and enable monitoring performance for algorithmic drift.

Data infrastructure is an important influencer for the adoption of AI innovations. AI requires a continuous supply of high-quality data. Data quality issues may entail accuracy, completeness, consistency, timeliness, integrity, relevance, data collection, preprocessing, management, data governance, and data labeling [47]. Storage challenges, processing challenges,
data management challenges, data heterogeneity, data privacy and security, bias and representativeness, and data access are also data quality considerations [47]. An appropriate data infrastructure, including its maintenance and evolution over time, is a prerequisite for successful AI applications.

Management of eye health necessitates a multidisciplinary team with a dynamic flow of information between treating doctors [48]. Holley and Lee’s [4] qualitative research found that PCPs had poor communication with eye care providers and the PCPs desire changes in the current referral-to-eye-care system. Better communication between PCPs and eye care professionals, further implementation of EHRs, and increasing eye screening in primary care clinics were common themes. Moudgil et al [48] found that 80% of the physicians communicated with ophthalmologists sometimes, whereas only 10% ensured communication at all times. The information sought by the treating physicians from the ophthalmologists regarding their referral for ocular findings included severity, the grading of DR, other ocular changes, need for intervention, and the frequency of screening and follow-up based on changes observed.

Finally, ethical considerations call for AI systems to adhere to the principles of fairness and nondiscrimination [49,50]. Advances in modern medicine are sometimes stymied by the inability to translate evidence-based care to all patients [51]. Transparency of AI models is essential to be able to evaluate and ensure their relevance for diverse populations and the ability to translate the innovations to all settings of care.

Limitations

Several limitations are inherent in the use of aggregated clinical data. Longitudinal data on patients are limited, and this, by extension, limits projects such as ours in their ability to predict the development of future pathology. Although the data set does derive information from EHRs, including Epic, Cerner, GE, and McKesson, the actual physicians’ notes are not available for analysis. Aggregated data also disproportionally represent hospital encounters and underrepresent outpatient visits [52]. Attempts to mitigate some of these deficiencies in the feature and outcome engineering methods are described before. A certain degree of circumspection should be exercised when applying this model more broadly to other databases that may have used different NLP protocols.

A challenge with deploying these models in their current form is that the richness of data (ie, number of parameters) to be input into the models must be balanced against the labor the clinician must expend entering them. The authors sought to reduce feature input without substantially affecting model predictive performance. The goal is to develop tools that will aid clinicians and reduce the number of undiagnosed serious ophthalmic conditions. Empirically based analyses such as those presented here are exploratory and intended to generate insights worthy of subsequent investigation with different study designs and methods that are better suited for causal inference.

It is important to note that data quality and representativeness are a potential issue for ML model training from EHRs and other clinical databases. EHR data can be incomplete, inconsistent, or erroneous, given the nature of the data collection and documentation. EHR data can also be biased toward populations with better access to health care. Some of these issues (eg, access) are inherent to our health care system in general and are not specific to EHR data. Regardless of the source of the issue, it is important to note that models trained and tested on EHR data may not be generalizable to the larger population.

Conclusion

In summary, this research demonstrates real patient triage potential by deploying AI strategies directly to PCP EHRs. In addition, based on the original data pool (more than 80 million patients), the final study population size (1,486,078 patients, 50% of whom were controls) and the 9 subpathologies using 5 different analytical modeling approaches, the authors believe this study to be one of the largest AI projects in ophthalmology.

Acknowledgments

This study was funded by Johnson & Johnson Vision, Inc. The sponsor participated in the design of the study, conducting the study, data collection, data management, data analysis, interpretation of the data, and preparation, review, and approval of the manuscript.

Conflicts of Interest

JAY is a consultant for Johnson & Johnson Vision, Inc. CWC, CWS, CEH and CAB are employees of Johnson & Johnson. SVM was a contractor with Johnson & Johnson at the time of the study.

Multimedia Appendix 1

Patient inclusion and exclusion criteria and attrition.

References


50. Weidner L, Fischer M. Role of ethics in developing AI-based applications in medicine: insights from expert interviews and discussion of implications. JMIR AI 2024 Jan 12;3:e51204. [doi: 10.2196/51204]


Abbreviations

AI: artificial intelligence
AMD: age-related macular degeneration
AUC: area under the curve
DEMGD: dry eye and meibomian gland dysfunction
DL: deep learning
DR: diabetic retinopathy
DT: decision tree
EHR: electronic health record
GLM: generalized linear model
ICD-10: International Classification of Diseases, 10th Revision
LR: logistic regression without regularization
L1-LR: L1-regularized logistic regression
ML: machine learning
NLP: natural language processing
NPDR: nonproliferative diabetic retinopathy
OR: odds ratio
OSD: ocular surface disease
PCP: primary care practitioner
PDR: proliferative diabetic retinopathy
RF: random forest
XGBoost: Extreme Gradient Boosting

©Joshua A Young, Chin-Wen Chang, Charles W Scales, Saurabh V Menon, Chantal E Holy, Caroline Adrienne Blackie. Originally published in JMIR AI (https://ai.jmir.org), 12.03.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR AI, is properly cited. The complete bibliographic information, a link to the original publication on https://www.ai.jmir.org/, as well as this copyright and license information must be included.
Corrigenda and Addenda

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

Max Rollwage¹, BSc, MSc, MPhil, PhD; Johanna Habicht¹, MSci; Keno Juechems¹, BSc, MSc, PhD; Ben Carrington¹, BSc; Sruthi Viswanathan¹, B’Tech, MRes; Mona Stylianou², BA, PGDip; Tobias U Hauser¹,³,⁴,⁵, PhD; Ross Harper¹, MA, MRes, PhD

¹Limbic Limited, London, United Kingdom
²Everyturn Mental Health, Gosforth, United Kingdom
³Max Planck University College London Centre for Computational Psychiatry and Ageing Research, University College London, London, United Kingdom
⁴Department of Psychiatry and Psychotherapy, Medical School and University Hospital, Eberhard Karls University of Tübingen, Tübingen, Germany
⁵German Center for Mental Health (DZPG), Tübingen, Germany

Corresponding Author:
Max Rollwage, BSc, MSc, MPhil, PhD
Limbic Limited
Kemp House
160 City Road
London,
United Kingdom
Phone: 44 07491263783
Email: max@limbic.ai

Related Article:
Correction of: https://ai.jmir.org/2023/1/e44358
doi:10.2196/57869

In “Using Conversational AI to Facilitate Mental Health Assessments and Improve Clinical Efficiency Within Psychotherapy Services: Real-World Observational Study” (JMIR AI 2023;2:e44358) the authors noted one error.

One author, Sruthi Viswanathan, was inadvertently omitted from the authorship list in the original publication of the paper. Sruthi Viswanathan has now been added to the authorship of the published paper as the fifth author, with the degrees “BTech, MRes” and the following affiliation:

Limbic Limited, London, United Kingdom

In accordance, the Conflict of Interest statement has also been updated to include this author. The originally published statement appeared as follows:

MR, KJ, JH, BC, and RH are employed by Limbic Limited and hold shares in the company. TUH works as a paid consultant for Limbic Limited and holds shares in the company.

This statement has been corrected to:

MR, KJ, JH, BC, SV and RH are employed by Limbic Limited and hold shares in the company. TUH works as a paid consultant for Limbic Limited and holds shares in the company.

The correction will appear in the online version of the paper on the JMIR Publications website on March 12, 2024, together with the publication of this correction notice. Because this was made after submission to PubMed, PubMed Central, and other full-text repositories, the corrected article has also been resubmitted to those repositories.

https://ai.jmir.org/2023/1/e44358

(JMIR AI 2024;3:e57869) doi:10.2196/57869