JMIR Medical Informatics

Contents

Reviews

Viewpoints

Original Papers

Review

State of the Art of Machine Learning–Enabled Clinical Decision Support in Intensive Care Units: Literature Review

Na Hong[1], PhD; Chun Liu[1], PhD; Jianwei Gao[1], PhD; Lin Han[1], MSc, MD; Fengxiang Chang[1], PhD; Mengchun Gong[1], MD, PhD; Longxiang Su[2], MD, PhD

[1]Digital Health China Technologies Ltd Co, Beijing, China

[2]Department of Critical Care Medicine, State Key Laboratory of Complex Severe and Rare Diseases, Peking Union Medical College Hospital, Chinese Academy of Medical Science and Peking Union Medical College, Beijing, China

Corresponding Author:
Longxiang Su, MD, PhD
Department of Critical Care Medicine
State Key Laboratory of Complex Severe and Rare Diseases, Peking Union Medical College Hospital
Chinese Academy of Medical Science and Peking Union Medical College
No.1 Shuaifuyuan Wangfujing Dongcheng District
Beijing
China
Phone: 86 10 69152308
Email: sulongxiang@vip.163.com

Abstract

Background: Modern clinical care in intensive care units is full of rich data, and machine learning has great potential to support clinical decision-making. The development of intelligent machine learning–based clinical decision support systems is facing great opportunities and challenges. Clinical decision support systems may directly help clinicians accurately diagnose, predict outcomes, identify risk events, or decide treatments at the point of care.

Objective: We aimed to review the research and application of machine learning–enabled clinical decision support studies in intensive care units to help clinicians, researchers, developers, and policy makers better understand the advantages and limitations of machine learning–supported diagnosis, outcome prediction, risk event identification, and intensive care unit point-of-care recommendations.

Methods: We searched papers published in the PubMed database between January 1980 and October 2020. We defined selection criteria to identify papers that focused on machine learning–enabled clinical decision support studies in intensive care units and reviewed the following aspects: research topics, study cohorts, machine learning models, analysis variables, and evaluation metrics.

Results: A total of 643 papers were collected, and using our selection criteria, 97 studies were found. Studies were categorized into 4 topics—monitoring, detection, and diagnosis (13/97, 13.4%), early identification of clinical events (32/97, 33.0%), outcome prediction and prognosis assessment (46/97, 47.6%), and treatment decision (6/97, 6.2%). Of the 97 papers, 82 (84.5%) studies used data from adult patients, 9 (9.3%) studies used data from pediatric patients, and 6 (6.2%) studies used data from neonates. We found that 65 (67.0%) studies used data from a single center, and 32 (33.0%) studies used a multicenter data set; 88 (90.7%) studies used supervised learning, 3 (3.1%) studies used unsupervised learning, and 6 (6.2%) studies used reinforcement learning. Clinical variable categories, starting with the most frequently used, were demographic (n=74), laboratory values (n=59), vital signs (n=55), scores (n=48), ventilation parameters (n=43), comorbidities (n=27), medications (n=18), outcome (n=14), fluid balance (n=13), nonmedicine therapy (n=10), symptoms (n=7), and medical history (n=4). The most frequently adopted evaluation metrics for clinical data modeling studies included area under the receiver operating characteristic curve (n=61), sensitivity (n=51), specificity (n=41), accuracy (n=29), and positive predictive value (n=23).

Conclusions: Early identification of clinical and outcome prediction and prognosis assessment contributed to approximately 80% of studies included in this review. Using new algorithms to solve intensive care unit clinical problems by developing reinforcement learning, active learning, and time-series analysis methods for clinical decision support will be greater development prospects in the future.

Introduction

With the popularization of electronic health records, medical equipment, and the improvement of detection methods, patient data are generated in large amounts every day in intensive care units. In traditional clinical data analysis, models and tools can only make use of a limited number of variables in clean and well-organized data. Machine learning has enabled clinical decision support research and applications to generate actionable insights, by utilizing large amounts of intensive care unit patient data, that are useful in many clinical scenarios.

Machine learning, sometimes called the data-driven method, uses statistical analysis models and computational technologies, allowing computer systems to learn from patient data and discover unknown clinical situations. Supervised learning, unsupervised learning, and reinforcement learning are the 3 main types of machine learning [1] used to predict or guide the treatment of patients who are critically ill.

In supervised machine learning tasks, a function maps an input to an output based on example input–output pairs. Functions are inferred from labeled training data. Classification and regression methods, which include but are not limited to linear regression, logistic regression, decision tree, random forest, and support vector machine, are common supervised learning methods.

In unsupervised machine learning tasks, patterns are learned from untagged data. Models are designed to identify or partition large data sets into subsections or clusters that share similar characteristics. In intensive care unit–related tasks, unsupervised learning enables the discovery of latent structures or patient subgroups in specific cohorts [2]. Commonly used unsupervised learning models include clustering, auto-encoding, and principal component analysis.

Reinforcement learning is concerned with how intelligent agents ought to take actions in an environment to maximize the notion of cumulative rewards. The environment is typically defined by a discrete-time stochastic control process called the Markov decision process. In an intensive care unit, clinicians often need to determine treatment plans and make clinical decisions. Reinforcement learning models have great potential for solving these types of problems by providing targeted treatment plans for each patient or patient status and assisting clinicians in making efficient decisions [3-8].

Although there are still challenges when data from multiple sources must be combined, and the performance and ability of machine learning is limited by the volume and quality of data, a number of clinical decision support studies [9,10] have demonstrated the ability to use sophisticated machine learning models to solve certain intensive care unit–related tasks, and their performance has been shown to be comparable with human abilities, and for certain tasks, even it potentially exceeds human abilities [7,11].

We sought to focus on machine learning research and applications adapted to clinical decision support in intensive care units, which may directly help clinicians diagnoses accurately, predict outcomes, identify risk events, or decide treatments at the intensive care unit point of care.

Methods

Search Strategy

We searched for papers in the PubMed database that had been published prior to October 2020 using a query combination of MeSH terms ("intensive care unit," "critical care," "machine learning," "artificial intelligence," "decision support systems, clinical") and keywords in the title or abstract keywords related to *machine learning* ("machine learning," "artificial intelligence," "prediction model," "predictive model," "predictive modeling," "artificial learning," "predictive analysis," "machine intelligence," "data driven," "data-driven," "statistical learning," "neural network," "deep learning," "reinforcement learning," "time series," "time-series," "algorithm"), *decision-making* ("clinical decision support system," "medical decision," "decision tool," "support tool," "clinical decision," "physician decision," "clinician decision," "decision algorithm," "CDSS," "CDS," "clinical management," "decision making," "decision-making"), and *intensive care units* ("intensive care," "ICU," "critical care," "intensive care unit").

Selection Criteria

We included English-language papers that reported studies (both prospective and retrospective studies) on clinical decision support, with machine learning methods that targeted a specific clinical scenario of intensive care units. We excluded papers that were systematic reviews and meta-analyses, studies of clinical decision support system implementations or clinical decision support system usability evaluations, studies that described rule-based clinical decision support system, studies that used data that were not from patients in intensive care units (eg, studies for intensive care unit admission prediction but using patient data from other departments, such as emergency or surgery departments), studies with outcomes irrelevant to regular intensive care unit clinical care (eg, studies about estimation of caffeine regimens), and studies that did not use machine learning methods (eg, studies using clinical scores or statistical analysis on small samples).

Data Analysis

We extracted the following information from selected papers for content analysis: study cohort, machine learning models, analysis variables, evaluation methods, and research topics.

Study Cohort

In general, the greater the number of data sets to which a machine learning model is applied, the stronger its generalization capabilities. Therefore, we investigated the inclusion cohorts and distribution centers of each study and classified these studies into single-site or multisite studies accordingly. We also

classified studies by c, the sample size of studies: $c<500$, $500<c<2000$, $2000<c<5000$, $5000<c<10,000$, $10,000<c<50,000$, and $c>50,000$.

Machine Learning Models

The model methods or algorithms used in each paper were reviewed for analysis, and model methods were categorized as supervised learning, unsupervised learning, or reinforcement learning.

We reviewed variables or features used for modeling in each study. According to routine intensive care unit practices, we classified these variables into 12 groups: demographic variables, vital signs, symptoms, laboratory values, ventilation parameters, medications, nonmedicine therapy, comorbidities, fluid balance, scores, medical history, and outcome. Given the wide range of variable expressions in papers, such as formal medical terms, abbreviations, acronyms, and capitalizations, variable name normalization was implemented using text processing and manual annotation methods. As some studies used self-defined features or derived data for their special study purpose, variables used in only 1 study were excluded.

Evaluation Methods

To determine the applicability and potential impact of various machine learning models for clinicians and patients (ie, in applications), model evaluation methods are important components of model development. We reviewed evaluation metrics used for measuring model performance.

Research Topics

In addition to overall quantitative analysis, which included all studies, selected papers were divided into 4 topics for detailed analysis: detection and monitoring for diagnosis, early identification of clinical events, patient outcome prediction, and treatment decisions.

Results

General

A total of 643 papers were found. The number of machine learning–enabled intensive care unit clinical decision support system research papers published in the PubMed database has been continuously increasing between January 1980 and October 2020 (Figure 1).

Among the 643 papers identified and assessed for eligibility, 14 non–English language papers, 55 clinical decision support system implementations and clinical decision support system usability evaluations, 114 reviews and meta-analyses, 35 expert system clinical decision support system studies, 68 studies not about intensive care unit clinical questions, 76 studies using patient data from other clinical departments or with outcomes irrelevant to regular intensive care unit clinical care, 107 studies that used methods other than machine learning, and 77 studies for which full-text papers were unavailable were excluded (Figure 2); therefore, 97 papers remained (Table 1).

Most studies used data from adult patients (n=82, 84.5%); however, 8 studies used data from pediatric patients (8.2%) and 7 studies used data from neonates (7.2%). Two-thirds of the studies (65/97, 67.0%) were developed from single-center data sets, and 32 (33.0%) were developed from a multicenter data set; cohort sizes also varied ($c<500$: 35/97, 36%; $500<c<2000$: 19/97, 20%; $2000<c<5000$: 12/97, 12%; $5000<c<10000$: 10/97, 10%; $10000<c<50000$: 16/97, 16%; $c>50,000$: 7/97, 7%).

The vast majority of studies used supervised learning (88/97, 91%), and only a few used unsupervised learning (3/97, 3%) or reinforcement learning (6/97, 6%). In total, 849 variables for model analysis were extracted. The most frequent variable categories are shown in Table 1, and the top 20 most frequently used variables are shown in Figure 3.

Most studies used more than 1 evaluation metric. The most frequently used were area under receiver operating characteristic curve (n=57), sensitivity (n=37), specificity (n=31), and accuracy (n=24).

XSL•FO
RenderX

Figure 1. Growth in number of publications.
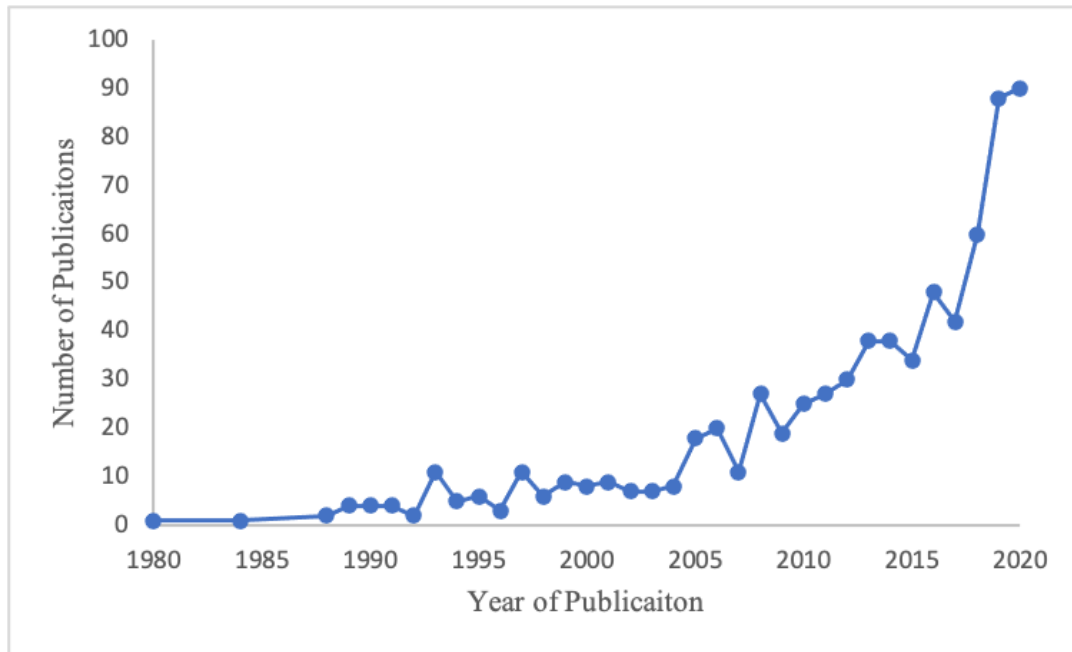
Figure 2. Article review process. CDSS: clinical decision support system; ICU: intensive care unit.
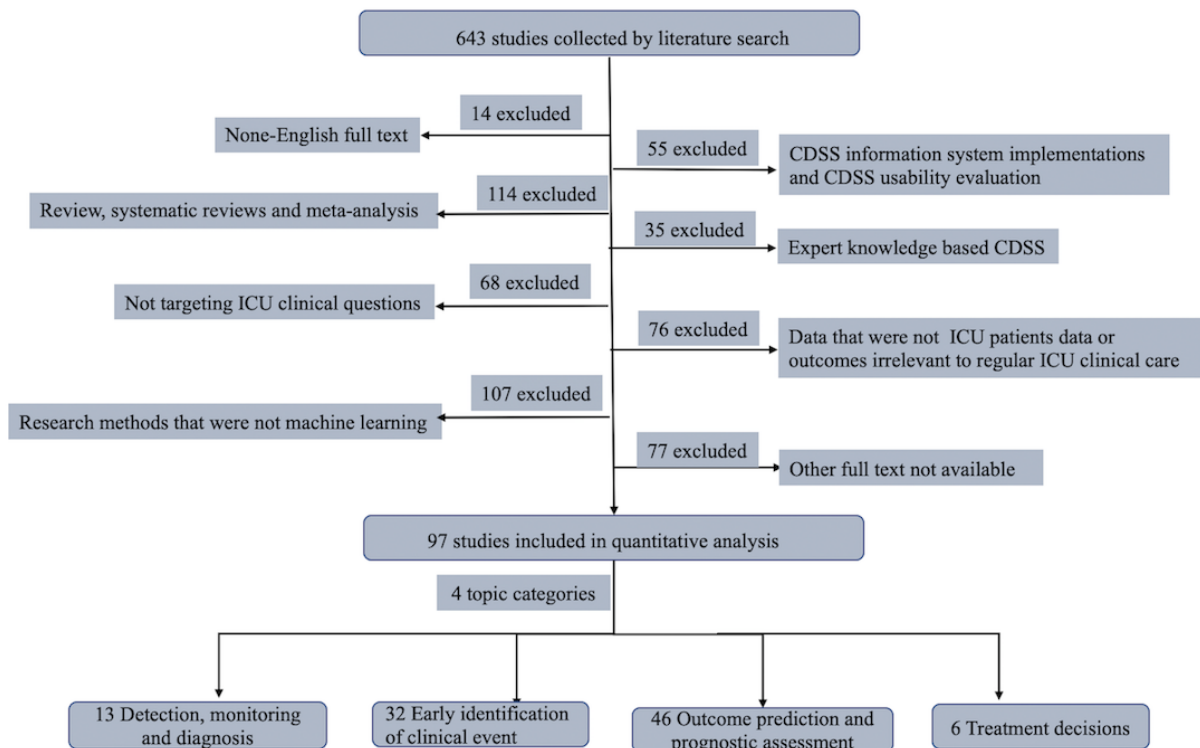
Table 1. General characteristics of the selected studies.

Characteristic	Value (n=97), n
Types of decision support	
Detection, monitoring, and diagnosis	13
Early identification of clinical events	32
Outcome prediction and prognostic assessment	46
Treatment decisions	6
Population	
Adult	82
Pediatric patients	8
Neonates	7
Medical setting	
Single-center	65
Multicenter	32
Type of machine learning	
Supervised learning	88
Unsupervised learning	3
Reinforcement learning	6
Type of variables	
Demographic variables	74
Laboratory values	59
Vital signs	55
Scores	48
Ventilation parameters	43
Comorbidities	27
Medications	18
Outcome	14
Fluid balance	13
Nonmedicine therapy	10
Symptoms	7
Medical history	4
Type of evaluation method, n[a]	
Area under the receiver operating characteristic curve	57
Sensitivity	37
Specificity	31
Accuracy	24
Positive predictive value	11

[a]More than 1 variable type could be used in each study.

XSL•FO
RenderX

Figure 3. Top 20 most frequently used variables. DBP: diastolic blood pressure; FiO2: fractional inspired oxygen; GCS: Glasgow Coma Scale; HR: heart rate; MBP: mean blood pressure; MV: mechanical ventilation; PaO2-partial pressure of oxygen; RR: respiratory rate; SBP: systolic blood pressure; SCR: creatine; SpO2: peripheral capillary oxygen saturation; WBC: white blood cell count.
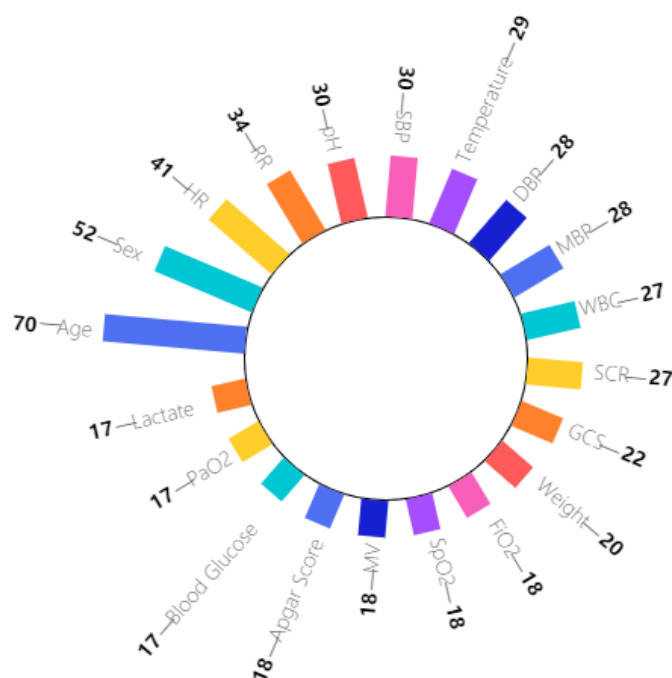
Monitoring, Detection, and Diagnosis

Overview

Among 13 studies, 4 (30.8%) studies [12-15] focused on monitoring or detection of physiological indicators, 3 studies (23.1%) [16-18] focused on the of mechanical ventilation abnormalities (in particular, patient-ventilator asynchrony), 4 studies (30.8%) [19-22] used electroencephalography (EEG) to diagnose brain diseases, and 2 studies (15.4%) [11,23] studies focused on infections. Variables used included demographic variables (n=5), vital signs (n=6), laboratory values (n=5), ventilation parameters (n=5), comorbidities (n=1), and outcome (n=1).

Most data were obtained from a single center (11/13, 84.6%), and only 2 studies (2/13, 15.4%) used multicenter data sets. Some studies (3/13, 23.1%) used data from public databases, such as the MIMIC database, the public NIH Chest-XRay14, and PLCO data sets (Multimedia Appendix 1).

The top 3 models used were neural network (n=4), tree (n=3), and random forest (n=3) models. Support vector machine models were used twice (n=2). Other models, such as logistic regression, and linear regression were only used in 1 study each.

Model performance was mainly evaluated with sensitivity (n=7), specificity (n=8), area under the receiver operating characteristic curve (n=3), and accuracy (n=3), whereas other evaluation methods such as equal error rates, F1 score, recall, and κ coefficients were each used only once.

Monitoring of Physiological Indicators

Quinn et al [13] provided a general model for inferring hidden factors from clinical data and was successfully applied to the major task of monitoring premature infants in the intensive care unit. Eshelman et al [12] described an algorithm consisting of

a set of rules for identifying intensive care unit patients who may become hemodynamically unstable. Taking into account the individual differences of intensive care unit patients, Zhang and Szolovits [15] developed an algorithm based on personalized vital signs data to improve the accuracy of alarms. Charbonnier [14] extracted online temporal episodes from the high-frequency physiological parameters of intensive care unit patients to visually support signal interpretation.

Mechanical Ventilation

Mechanical ventilation is widely used in intensive care units, during which a series of parameters need to be monitored. Kwok et al [16] established a nonlinear adaptive neuro-fuzzy inference system model for fractional inspired oxygen estimation, which reduced the need for invasive inspections. Two groups of researchers discussed the problem of patient-ventilator asynchrony, and developed a classifier based on machine learning to detect abnormal waveforms [17,18].

Electroencephalography Monitoring

EEG monitoring plays an important role in the detection of brain function and the diagnosis of brain disease. Koolen et al [19] developed a method for the automated classification of neonatal sleep states via EEG. Golmohammadi et al [21] presented a system that can achieve high-performance classification of EEG events that might correlate with epilepsy, metabolic encephalopathy, cerebral hypoxia, and ischemia. Farzaneh et al [20] developed a machine learning framework to automatically segment and assess the severity of patients with subdural hematoma during traumatic brain injuries [20].

Diagnosis of Infection

Infections are an important clinical issue in intensive care. Sepsis is a common and serious condition in the intensive care unit that results from an overreaction to infection that damages

tissues and organs and can lead to complications, making it one of the leading causes of hospital-related deaths [24]. A high-performance algorithm, InSight, was demonstrated to be superior to the commonly used Modified Early Warning Score, Simplified Acute Physiology Score, and Systemic Inflammatory Response Syndrome score for the diagnosis of patients with alcohol use disorder combined with sepsis shock [23]. In addition, it is still challenging to explain lung opacity in radiography of the supine chest of patients with lung infection in the intensive care unit—Rueckel et al [11] evaluated a prototype artificial intelligence algorithm that could classify underlying lung opacity, which might suggest a diagnosis pneumonia.

Early Identification or Prediction of Clinical Events

Overview

Clinical event prediction, the use of data from electronic health records to predict the occurrence of certain events or the best time to give treatment, is one of the most important aspects of intensive care unit clinical decision support system. Among 32 clinical event prediction studies, 3 (9.4%) were related to acute kidney injury, 11 (34.4%) were related to infection prediction, 8 (25%) were related to respiratory diseases, and 10 (31.3%) were related to other predictions and evaluations (Multimedia Appendix 1).

In intensive care unit clinical prediction and evaluation studies, up to 87 variables were used in a single paper. Categories of variables, in order of frequency, were laboratory values (n=25), demographic variables (n=25), vital signs (n=20), scores (n=18), ventilation parameters (n=14), fluid balance (n=8), medications (n=7), comorbidities (n=7), outcome (n=4), nonmedicine therapy (n=3), symptoms (n=3), and medical history (n=1).

More than three-quarters of the studies (25/32, 78%) were based on data from a single center, 10 of which were from the freely available public database Medical Information Mart for Intensive Care II or III. Multi-institutional data were used in the other studies (7/32, 22%).

Logistic regression was the most commonly used method (11/32, 34%), followed by neural networks (7/32, 21%), and random forest (6, 19%). Support vector machine and decision tree models were each used in 5 (15.6%) studies. Naive Bayes, gradient boosting tree model, extreme gradient boosting, fuzzy model, and Insight each appeared twice (6.3%).

Sensitivity (n=16) and area under receiver operating characteristic curve (n=17) were the most commonly used evaluation metrics, followed by specificity (n=12) and accuracy (n=12). The following metrics appeared in fewer than 10 papers: positive predictive value (n=3), F1 score (n=4), and mean absolute error (n=2).

Acute Kidney Injury Prediction

Early prediction of acute kidney injury has a high value for the long-term survival and quality of life of critically ill patients. Acute kidney injury is often associated with high morbidity and mortality rates in intensive care units. The status of other vital organs, initiation of therapy, patient response, and preexisting comorbidities can all contribute to the development of acute

kidney injury [25]. Multiple machine learning methods have been utilized and compared to analyze unstructured clinical records and structured physiological measurements to identify early episodes of acute kidney injury [26]. Soliman et al [25] studied the prognostic impact of early acute kidney injury predicted by data from the first day of admission. One study [27] focused on patients younger than 21 years, who are more likely to recover from disease.

Prediction of Sepsis and Infection

Early identification and treatment is the key to survival for many sepsis and infection patients [28], but it is difficult for clinicians to predict before it occurs, because it is extremely complex and each patient is different. Early prediction of sepsis using interpretable or uninterpretable machine learning models can help clinicians enhance the accuracy of fever workup [28] to identify and intervene in a timely manner [29-33]. One research aim is to make accurate predictions with as little electronic health record data as possible [34]. Mao et al achieved early prediction of sepsis using only vital signs validated in multiple centers [35]. The prediction of neonatal sepsis has also received substantial research attention in recent years [36,37]. One paper [38] focuses on predicting infections caused by a specific microorganism—invasive fungal disease due to *Candida* species—in intensive care unit patients.

Prediction of Respiratory Disease and Mechanical Ventilation

Respiratory management in the intensive care unit is an important aspect of critical care and treatment. Early diagnosis of respiratory critical illness has a significant impact on patient prognosis [39]. In addition, maintenance of cardiopulmonary function is required in patients admitted to the intensive care unit due to acute symptoms such as direct trauma, pulmonary infection, heart failure, and sepsis. Machine learning methods can help predict the onset of acute respiratory disease in patients, especially in pediatric patients. Sauthier et al [40] used random forest and logistic regression to predict the time of acute hypoxic respiratory failure in critically ill children with severe influenza. Messinger et al [39] applied a cascaded artificial neural network to design new respiratory scores for early identification of asthma in young children. In addition, early prediction of acute respiratory distress syndrome was studied because of its high morbidity and mortality [41].

Furthermore, ventilator weaning and reintubation after weaning are currently well studied [42,43] in intensive care unit clinical decision support system literature, as well as the effect of drugs on intubation [44]. Moreover, predicting patient oxygen saturation after ventilation [45] and risk factors for failure of mechanical ventilation [46] can help health care professionals respond in a time manner.

Other Predictions and Evaluations

There were 10 papers that could not be classified; we simply put them into one class separately. There were forecasts for detection and monitoring indicators, such as urine output after fluid administration [47], glucose [48], lactic acid [49], and activated partial thromboplastin time [50]. Lin [47] established a gradient tree-based machine learning model implemented with

extreme gradient boosting algorithms to predict urine output in sepsis patients after fluid resuscitation to prevent fluid overload-related complications. Pappada et al [48,49] developed a neural network–based model to obtain a complete trajectory of glucose values up to 135 minutes in advance. Mamandipoor et al [49] combined least absolute shrinkage and selection operator regression, random forest, and long short-term memory to predict blood lactate concentration in patients in the intensive care unit. Our previous study also compared multiple machine learning approaches to guide clinical heparin administration by predicting the range of activated partial thromboplastin time values [50]. There were also studies that aimed to reduce unnecessary laboratory tests to streamline the process and reduce the burden on patients [51,52]. Predicted clinical events also included acute traumatic coagulopathy [53], delirium [54], advanced anemia [55], and fluid resuscitation therapy [56].

Outcome Evaluation and Prognostic Assessment

Overview

Of 46 papers that used machine learning for outcome evaluation for patients who were critically ill, 11 papers (23.9%) predicted overall mortality and survival, 23 papers (50%) predicted the outcomes of patients with certain diseases, and 12 papers (26.1%) included treatment prognosis, length of stay in the intensive care unit, and other outcome evaluations (Multimedia Appendix 1).

Categories of variables, in order of frequency, were demographic variables (n=39), scores (n=24), laboratory values (n=23), ventilation parameters (n=20), vital signs (n=18), comorbidities (n=17), medications (n=10), outcome (n=8), nonmedicine therapy (n=7), fluid balance (n=4), symptoms (n=4), and medical history (n=3).

Of the 46 outcome prediction studies, 25 (54.3%) were based on single-center data, 6 of which used data from MIMIC II and III, and the other 21 studies (45.7%) made use of multicenter data.

Logistic regression was the most commonly used method (27/46, 59%), followed by random forest (9/46, 20%), random forest (8/46, 17%), support vector machine (7/46, 15.2%) and decision tree model (5/46, 11%) studies. The gradient boosting tree model appeared in 4 (9%) studies, and adaptive boosting and linear regression each appeared twice (4.3%). Other models that appeared only once are not discussed here.

Area under receiver operating characteristic curve (n=37) was the evaluation metric used most often, followed by sensitivity (n=14), specificity (n=11), positive predictive value (n=4), accuracy (n=8), negative predictive value (n=6),F1 score (n=2), Matthews correlation coefficient (n=2), and Brier score (n=2).

Overall Intensive Care Unit Patient Outcomes

Typical outcomes were overall mortality [57-62], survival [63], and long-term quality of life [64]. Mortality [65,66] and survival status at 1 year [67] in critically ill patients aged 80 years and older were also studied using machine learning methods.

Outcomes of Patients With Specific Diseases

Patients with sepsis and infection remain one of the most studied populations in terms of mortality (generally 28 days) [68-72], followed by acute kidney injury [72-75]. There is an increasing trend in outcome prediction studies in critically ill patients with liver disease—acute liver injury [76,77], cirrhosis [77], and advanced liver disease [78] have been studied using machine learning. In patients with severe cancer, 30- [79] and 120-day [80] survival rates were studied retrospectively with logistic regression models.

For cardiac disease, Lee et al [81] used EEG data to predict the outcome of children with cardiac arrest and Murtuza et al [82] found that arterial blood lactate levels can be associated with mortality in children who have undergone cardiac surgery. For brain diseases, the outcomes of patients with subarachnoid hemorrhage [83] and severe traumatic brain injury [84] have been analyzed. Wildman et al [85] predicted the impact of chronic obstructive pulmonary disease and asthma on mortality in critically ill patients. Daly et al [86] used logistic regression to study the relationship between early discharge and mortality with the intention of reducing mortality in this group of intensive care unit patients. Other papers [87-89] examined patient outcomes and factors influencing them after deterioration. Ebadollahi et al [90] predicted the temporal trajectory of physiological data with patient similarity, with the aim to identify universal patterns of disease progression from a large amount of clinical practice data, to establish a generalized computer-aided clinical decision support framework for personalized treatment.

Treatment Prognosis and Intensive Care Unit Stay Time Evaluation

Evaluating the outcome of certain treatments through machine learning can help medical professionals refine their treatments to achieve better therapeutic effects. Evaluation of outcomes after extubation based on continuous vital sign information and static characteristics of children can help adjust the timing of extubation to reduce mortality [91-93]. Evaluation of prolonged mechanical ventilation [94] and 1-year and 5-year functional survival [95] after cardiac surgery was used to help adjust and optimize postsurgical care practices. Evaluating the length of stay in the intensive care unit [96,97] and the risk of readmission after discharge from the intensive care unit [98] to effectively forecast the trend of the disease could improve treatment and care. In addition, designing and improving critical illness scores to indicate disease severity [99-101] was studied. For example, McRae et al [102] designed a score to quickly determine the severity of COVID-19 and achieved optimistic results in 160 individuals.

Treatment Decisions

Treatments, clinical determination, and decision-making in the intensive care unit were studied in 6 papers [3-8]. These papers focused on various clinical questions and mainly used a reinforcement learning model. Among them, 4 papers [3,5,7,8] (67%) addressed drug dosage, such as optimal vasopressin dose [3,7], heparin dosage [5], and morphine dosage [8]. The other

2 papers [4,6] (33%) studied the timing of mechanical ventilation extubation.

Categories of variables, in order of frequency, were vital signs (n=6), demographic variables (n=5), laboratory values (n=5), ventilation parameters (n=3), medications (n=4), fluid balance (n=2), scores (n=4), and comorbidities (n=1) (Multimedia Appendix 1).

Reinforcement learning models can be divided into conventional reinforcement learning models (that is, wherein the reward function is known and we only need to find a policy to maximize the reward function) and inverse reinforcement learning models (that is, wherein the reward function is unknown, and we have to learn the most reasonable reward function through the decision-making examples of clinicians)—4 papers used typical reinforcement learning model, and 2 papers used inverse reinforcement learning models.

All 6 papers used patient data from the intensive care units in US hospitals. Most papers used single-center data from MIMIC II (n=1) or MIMIC III (n=4), with c ranging from 707 to 96,156 (mean 22,256; median 7852).

Because the output of a reinforcement learning model is a policy that is not easy to evaluate, in these studies, the policy given by the model was compared with that actually given by the doctor; when the 2 policies differed, the effect of the reinforcement learning model was analyzed according to the actual clinical problem.

Discussion

From reviewed studies, we concluded that early identification of clinical outcome prediction and prognosis assessment contributed to approximately 80% of studies, and machine learning–based clinical decision support applications in intensive care unit could support timely bedside decision-making [15], transform data into more actionable insights or evidence-based clinical rules [101], assist disease diagnosis [30], predict adverse outcomes before they happen [76], enable continuous assessment of patient responses to critical care interventions [91], allow better management of highly complex situations and the best treatment decisions [3], ultimately reduce clinicians burden [52], and allow clinicians to have more time to deliver their knowledge, experience, and human care in practice [64].

We found that 91% (88/97) of reviewed studies used supervised learning methods. Unsupervised learning is commonly used for phenotyping or patient subgrouping [2], usually to discover new knowledge; therefore, explaining and validating subgroups or patterns with reasonable clinical meaning is a challenge. Reinforcement learning models have great potential for solving medical decision problems; however, to the best of our knowledge, there is a lack of sophisticated reinforcement learning models to guide intensive care unit decision-making [5]. Data-driven decision support tools will permit clinicians to function more efficiently, caring for more patients more safely; however the selection of a model should be tailored to the clinical scenario [9,10]; therefore, we need a better understanding of which algorithms are a best fit for which clinical scenarios.

We also found that many machine learning–based clinical prediction tasks are still challenging. First, not all the data collected from intensive care unit are good quality data or complete [7], particularly when data from different sources were included in one predictive model. Various data in the intensive care unit include general available data in the electronic health record, such as patient information, encounter information, diagnoses, intervention, routine laboratory data, imaging, natural language and physiologic data, as well as limited available information in the intensive care unit, such as social information, omics data, pathology, radiology, and wearable data [103]. This makes data preprocessing a difficult and time-consuming task. Second, parameter optimization was used to obtain the best parameter combination to improve model accuracy. Model parameters need to be determined and fitted using the training data set, and many adjustable hyperparameters must be tuned to obtain a model with optimal performance [104]. Generally, the more complex the model, the more parameters need to be adjusted, and the more difficult it is to adjust the parameters. For example, in logistic regression [74], usually only the regularization coefficient is adjusted; and in random forest models [53], the hyperparameters that need to be adjusted include the number of trees, the maximum depth of the tree, and the split criteria. Third, typically, the more complex the model, the higher the required sample size [105]. If the sample size is insufficient, overfitting occurs easily, which leads to instability or inaccuracy of the model. In some clinical scenarios, owing to the limited sample size, the use of complex models is limited [59]. Last, after developing the model, prospective evaluation using external data sets and clinical trials should be conducted before using the model in practice [106] to improve confidence in machine learning predictions [7]; however, performing strong validation of a machine learning model's generalizability and interpretability is challenging; internal validation approaches, such as cross-validation and bootstrapping, cannot guarantee the quality of a machine learning model due to potentially biased training data and the complexity of the validation procedure itself [107]. Lack of technical and semantic interoperability makes harmonization of patient data from one center to another costly. As inconsistent model results may be derived when adapting to new data sets [108], retraining models using data from other sources would minimize the cost and allow models to incorporate new clinical settings.

Future research should expand the innovation and exploration using new algorithms to solve intensive care unit clinical problems by developing reinforcement learning, active learning, and time-series analysis methods for clinical decision support. In addition, machine learning modeling requires recognition, understanding, and trust from intensive care unit clinicians. Model developers must provide full explanations of modeling methods, input, output, experimental and trial settings, clinical scenarios, and operation methods to clinicians. With the basis to understand, operate, and debug the outputs of a model, clinicians can have more confidence in accepting the model results and take action on the basis of that model's recommendations.

Acknowledgments

Authors' Contributions

LS and NH were responsible for study design and conception. NH and CL performed the search. NH, CL, JG, and LH were responsible for literature review and data analysis. NH, CL, JG, LH, MG and LS interpreted the results. FC supported data processing and analysis. All authors drafted and revised the manuscript for important intellectual content.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Supplementary information.
[DOCX File , 65 KB - medinform_v10i3e28781_app1.docx]

References

1. Pierre L. An introduction to machine learning. Language Technology Group (LTG). 2015. URL: https://studylib.net/doc/11539838/an-introduction-to-machine-learning-pierre-lison--languag [accessed 2022-02-03]

2. Su L, Zhang Z, Zheng F, Pan P, Hong N, Liu C, et al. Five novel clinical phenotypes for critically ill patients with mechanical ventilation in intensive care units: a retrospective and multi database study. Respir Res 2020 Dec 10;21(1):325 [FREE Full text] [doi: 10.1186/s12931-020-01588-6] [Medline: 33302940]

3. Srinivasan S, Doshi-Velez F. Interpretable batch IRL to extract clinician goals in ICU hypotension management. AMIA Jt Summits Transl Sci Proc 2020;2020:636-645 [FREE Full text] [Medline: 32477686]

4. Yu C, Liu J, Zhao H. Inverse reinforcement learning for intelligent mechanical ventilation and sedative dosing in intensive care units. BMC Med Inform Decis Mak 2019 Apr 09;19(Suppl 2):57 [FREE Full text] [doi: 10.1186/s12911-019-0763-6] [Medline: 30961594]

5. Nemati S, Ghassemi MM, Clifford GD. Optimal medication dosing from suboptimal clinical examples: a deep reinforcement learning approach. Annu Int Conf IEEE Eng Med Biol Soc 2016 Aug;2016:2978-2981. [doi: 10.1109/EMBC.2016.7591355] [Medline: 28268938]

6. Yu C, Ren G, Dong Y. Supervised-actor-critic reinforcement learning for intelligent mechanical ventilation and sedative dosing in intensive care units. BMC Med Inform Decis Mak 2020 Jul 09;20(Suppl 3):124 [FREE Full text] [doi: 10.1186/s12911-020-1120-5] [Medline: 32646412]

7. Komorowski M, Celi LA, Badawi O, Gordon AC, Faisal AA. The Artificial Intelligence Clinician learns optimal treatment strategies for sepsis in intensive care. Nat Med 2018 Nov;24(11):1716-1720. [doi: 10.1038/s41591-018-0213-5] [Medline: 30349085]

8. Lopez-Martinez D, Eschenfeldt P, Ostvar S, Ingram M, Hur C, Picard R. Deep reinforcement learning for optimal critical care pain management with morphine using dueling double-deep Q networks. Annu Int Conf IEEE Eng Med Biol Soc 2019 Jul;2019:3960-3963. [doi: 10.1109/EMBC.2019.8857295] [Medline: 31946739]

9. Greco M, Caruso PF, Cecconi M. Artificial intelligence in the intensive care unit. Semin Respir Crit Care Med 2021 Feb;42(1):2-9. [doi: 10.1055/s-0040-1719037] [Medline: 33152770]

10. Hanson CW, Marshall BE. Artificial intelligence applications in the intensive care unit. Crit Care Med 2001 Feb;29(2):427-435. [doi: 10.1097/00003246-200102000-00038] [Medline: 11269246]

11. Rueckel J, Kunz WG, Hoppe BF, Patzig M, Notohamiprodjo M, Meinel FG, et al. Artificial intelligence algorithm detecting lung infection in supine chest radiographs of critically ill patients with a diagnostic accuracy similar to board-certified radiologists. Crit Care Med 2020 Jul;48(7):e574-e583. [doi: 10.1097/CCM.0000000000004397] [Medline: 32433121]

12. Eshelman LJ, Lee KP, Frassica JJ, Zong W, Nielsen L, Saeed M. Development and evaluation of predictive alerts for hemodynamic instability in ICU patients. AMIA Annu Symp Proc 2008 Nov 06:379-383 [FREE Full text] [Medline: 18999006]

13. Quinn JA, Williams CKI, McIntosh N. Factorial switching linear dynamical systems applied to physiological condition monitoring. IEEE Trans Pattern Anal Mach Intell 2009 Sep;31(9):1537-1551. [doi: 10.1109/TPAMI.2008.191] [Medline: 19574617]

14. Charbonnier S. On line extraction of temporal episodes from ICU high-frequency data: a visual support for signal interpretation. Comput Methods Programs Biomed 2005 May;78(2):115-132. [doi: 10.1016/j.cmpb.2005.01.003] [Medline: 15848267]

XSL·FO

RenderX

15. Zhang Y, Szolovits P. Patient-specific learning in real time for adaptive monitoring in critical care. J Biomed Inform 2008 Jun;41(3):452-460 [FREE Full text] [doi: 10.1016/j.jbi.2008.03.011] [Medline: 18463000]

16. Kwok HF, Linkens DA, Mahfouf M, Mills GH. Adaptive ventilator FiO2 advisor: use of non-invasive estimations of shunt. Artif Intell Med 2004 Nov;32(3):157-169. [doi: 10.1016/j.artmed.2004.02.005] [Medline: 15531148]

17. Rehm GB, Han J, Kuhn BT, Delplanque J, Anderson NR, Adams JY, et al. Creation of a robust and generalizable machine learning classifier for patient ventilator asynchrony. Methods Inf Med 2018 Sep;57(4):208-219. [doi: 10.3414/ME17-02-0012] [Medline: 30919393]

18. Gholami B, Phan TS, Haddad WM, Cason A, Mullis J, Price L, et al. Replicating human expertise of mechanical ventilation waveform analysis in detecting patient-ventilator cycling asynchrony using machine learning. Comput Biol Med 2018 Jun 01;97:137-144. [doi: 10.1016/j.compbiomed.2018.04.016] [Medline: 29729488]

19. Koolen N, Oberdorfer L, Rona Z, Giordano V, Werther T, Klebermass-Schrehof K, et al. Automated classification of neonatal sleep states using EEG. Clin Neurophysiol 2017 Jun;128(6):1100-1108. [doi: 10.1016/j.clinph.2017.02.025] [Medline: 28359652]

20. Farzaneh N, Williamson CA, Jiang C, Srinivasan A, Bapuraj JR, Gryak J, et al. Automated segmentation and severity analysis of subdural hematoma for patients with traumatic brain injuries. Diagnostics 2020;10(10):773 [FREE Full text] [doi: 10.3390/diagnostics10100773] [Medline: 33007929]

21. Golmohammadi M, Harati Nejad Torbati AH, Lopez de Diego S, Obeid I, Picone J. Automatic analysis of EEGs using big data and hybrid deep learning architectures. Front Hum Neurosci 2019;13:76 [FREE Full text] [doi: 10.3389/fnhum.2019.00076] [Medline: 30914936]

22. Sorani MD, Hemphill JC, Morabito D, Rosenthal G, Manley GT. New approaches to physiological informatics in neurocritical care. Neurocrit Care 2007;7(1):45-52. [doi: 10.1007/s12028-007-0043-7] [Medline: 17565451]

23. Calvert J, Desautels T, Chettipally U, Barton C, Hoffman J, Jay M, et al. High-performance detection and early prediction of septic shock for alcohol-use disorder patients. Ann Med Surg (Lond) 2016 Jun;8:50-55 [FREE Full text] [doi: 10.1016/j.amsu.2016.04.023] [Medline: 27489621]

24. Timsit J, Perner A. Sepsis: find me, manage me, and stop me!. Intensive Care Med 2016 Dec 24;42(12):1851-1853. [doi: 10.1007/s00134-016-4603-1] [Medline: 27778045]

25. Soliman IW, Frencken JF, Peelen LM, Slooter AJC, Cremer OL, van Delden JJ, et al. The predictive value of early acute kidney injury for long-term survival and quality of life of critically ill patients. Crit Care 2016 Aug 03;20(1):242 [FREE Full text] [doi: 10.1186/s13054-016-1416-0] [Medline: 27488839]

26. Sun M, Baron J, Dighe A, Szolovits P, Wunderink RG, Isakova T, et al. Early prediction of acute kidney injury in critical care setting using clinical notes and structured multivariate physiological measurements. Stud Health Technol Inform 2019 Aug 21;264:368-372. [doi: 10.3233/SHTI190245] [Medline: 31437947]

27. Sanchez-Pinto LN, Khemani RG. Development of a prediction model of early acute kidney injury in critically ill children using electronic health record data. Pediatr Crit Care Med 2016 Jun;17(6):508-515. [doi: 10.1097/PCC.0000000000000750] [Medline: 27124567]

28. Fadlalla AMA, Golob JF, Claridge JA. Enhancing the fever workup utilizing a multi-technique modeling approach to diagnose infections more accurately. Surg Infect (Larchmt) 2012 Apr;13(2):93-101 [FREE Full text] [doi: 10.1089/sur.2008.057] [Medline: 20666579]

29. Nemati S, Holder A, Razmi F, Stanley MD, Clifford GD, Buchman TG. An interpretable machine learning model for accurate prediction of sepsis in the ICU. Crit Care Med 2018 Apr;46(4):547-553. [doi: 10.1097/CCM.0000000000002936] [Medline: 29286945]

30. Kam HJ, Kim HY. Learning representations for the early detection of sepsis with deep neural networks. Comput Biol Med 2017 Dec 01;89:248-255. [doi: 10.1016/j.compbiomed.2017.08.015] [Medline: 28843829]

31. Kaji DA, Zech JR, Kim JS, Cho SK, Dangayach NS, Costa AB, et al. An attention based deep learning model of clinical events in the intensive care unit. PLoS One 2019;14(2):e0211057 [FREE Full text] [doi: 10.1371/journal.pone.0211057] [Medline: 30759094]

32. Scherpf M, Gräßer F, Malberg H, Zaunseder S. Predicting sepsis with a recurrent neural network using the MIMIC III database. Comput Biol Med 2019 Oct;113:103395. [doi: 10.1016/j.compbiomed.2019.103395] [Medline: 31480008]

33. Wang S, Wu F, Wang B. Prediction of severe sepsis using SVM model. Adv Exp Med Biol 2010;680:75-81. [doi: 10.1007/978-1-4419-5913-3_9] [Medline: 20865488]

34. Desautels T, Calvert J, Hoffman J, Jay M, Kerem Y, Shieh L, et al. Prediction of sepsis in the intensive care unit with minimal electronic health record data: a machine learning approach. JMIR Med Inform 2016 Sep 30;4(3):e28 [FREE Full text] [doi: 10.2196/medinform.5909] [Medline: 27694098]

35. Mao Q, Jay M, Hoffman JL, Calvert J, Barton C, Shimabukuro D, et al. Multicentre validation of a sepsis prediction algorithm using only vital sign data in the emergency department, general ward and ICU. BMJ Open 2018 Jan 26;8(1):e017833 [FREE Full text] [doi: 10.1136/bmjopen-2017-017833] [Medline: 29374661]

36. Metsvaht T, Pisarev H, Ilmoja M, Parm U, Maipuu L, Merila M, et al. Clinical parameters predicting failure of empirical antibacterial therapy in early onset neonatal sepsis, identified by classification and regression tree analysis. BMC Pediatr 2009 Nov 24;9:72 [FREE Full text] [doi: 10.1186/1471-2431-9-72] [Medline: 19930706]

37. Mani S, Ozdas A, Aliferis C, Varol HA, Chen Q, Carnevale R, et al. Medical decision support using machine learning for early detection of late-onset neonatal sepsis. J Am Med Inform Assoc 2014 Mar;21(2):326-336 [FREE Full text] [doi: 10.1136/amiajnl-2013-001854] [Medline: 24043317]

38. Shahin J, Allen EJ, Patel K, Muskett H, Harvey SE, Edgeworth J, FIRE Study Investigators. Predicting invasive fungal disease due to Candida species in non-neutropenic, critically ill, adult patients in United Kingdom critical care units. BMC Infect Dis 2016 Sep 09;16:480 [FREE Full text] [doi: 10.1186/s12879-016-1803-9] [Medline: 27612566]

39. Messinger AI, Bui N, Wagner BD, Szefler SJ, Vu T, Deterding RR. Novel pediatric-automated respiratory score using physiologic data and machine learning in asthma. Pediatr Pulmonol 2019 Aug;54(8):1149-1155 [FREE Full text] [doi: 10.1002/ppul.24342] [Medline: 31006993]

40. Sauthier MS, Jouvet PA, Newhams MM, Randolph AG. Machine learning predicts prolonged acute hypoxemic respiratory failure in pediatric severe influenza. Crit Care Explor 2020 Aug;2(8):e0175 [FREE Full text] [doi: 10.1097/CCE.0000000000000175] [Medline: 32832912]

41. Le S, Pellegrini E, Green-Saxena A, Summers C, Hoffman J, Calvert J, et al. Supervised machine learning for the early prediction of acute respiratory distress syndrome (ARDS). J Crit Care 2020 Dec;60:96-102 [FREE Full text] [doi: 10.1016/j.jcrc.2020.07.019] [Medline: 32777759]

42. Hsu J, Chen Y, Chung W, Tan T, Chen T, Chiang JY. Clinical verification of a clinical decision support system for ventilator weaning. Biomed Eng Online 2013;12 Suppl 1:S4 [FREE Full text] [doi: 10.1186/1475-925X-12-S1-S4] [Medline: 24565021]

43. Miu T, Joffe AM, Yanez ND, Khandelwal N, Dagal AH, Deem S, et al. Predictors of reintubation in critically ill patients. Respir Care 2014 Feb;59(2):178-185 [FREE Full text] [doi: 10.4187/respcare.02527] [Medline: 23882103]

44. Isbister GK, Duffull SB. Quetiapine overdose: predicting intubation, duration of ventilation, cardiac monitoring and the effect of activated charcoal. Int Clin Psychopharmacol 2009 Jul;24(4):174-180. [doi: 10.1097/YIC.0b013e32832bb078] [Medline: 19494786]

45. Ghazal S, Sauthier M, Brossier D, Bouachir W, Jouvet PA, Noumeir R. Using machine learning models to predict oxygen saturation following ventilator support adjustment in critically ill children: A single center pilot study. PLoS One 2019;14(2):e0198921 [FREE Full text] [doi: 10.1371/journal.pone.0198921] [Medline: 30785881]

46. Rodríguez A, Ferri C, Martin-Loeches I, Díaz E, Masclans JR, Gordo F, Grupo Español de Trabajo Gripe A Grave (GETGAG)/Sociedad Española de Medicina Intensiva, Crítica y Unidades Coronarias (SEMICYUC) Working Group, 2009-2015 H1N1 SEMICYUC Working Group investigators. Risk factors for noninvasive ventilation failure in critically ill subjects with confirmed influenza infection. Respir Care 2017 Oct 11;62(10):1307-1315 [FREE Full text] [doi: 10.4187/respcare.05481] [Medline: 28698265]

47. Lin P, Huang H, Komorowski M, Lin W, Chang C, Chen K, et al. A machine learning approach for predicting urine output after fluid administration. Comput Methods Programs Biomed 2019 Aug;177:155-159. [doi: 10.1016/j.cmpb.2019.05.009] [Medline: 31319943]

48. Pappada SM, Owais MH, Cameron BD, Jaume JC, Mavarez-Martinez A, Tripathi RS, et al. An artificial neural network-based predictive model to support optimization of inpatient glycemic control. Diabetes Technol Ther 2020 May;22(5):383-394. [doi: 10.1089/dia.2019.0252] [Medline: 31687844]

49. Mamandipoor B, Majd M, Moz M, Osmani V. Blood lactate concentration prediction in critical care. Stud Health Technol Inform 2020 Jun 16;270:73-77. [doi: 10.3233/SHTI200125] [Medline: 32570349]

50. Su L, Liu C, Li D, He J, Zheng F, Jiang H, et al. Toward optimal heparin dosing by comparing multiple machine learning methods: retrospective study. JMIR Med Inform 2020 Jun 22;8(6):e17648 [FREE Full text] [doi: 10.2196/17648] [Medline: 32568089]

51. Yu L, Zhang Q, Bernstam EV, Jiang X. Predict or draw blood: an integrated method to reduce lab tests. J Biomed Inform 2020 Apr;104:103394. [doi: 10.1016/j.jbi.2020.103394] [Medline: 32113004]

52. Cismondi F, Celi LA, Fialho AS, Vieira SM, Reti SR, Sousa JMC, et al. Reducing unnecessary lab testing in the ICU with artificial intelligence. Int J Med Inform 2013 May;82(5):345-358 [FREE Full text] [doi: 10.1016/j.ijmedinf.2012.11.017] [Medline: 23273628]

53. Li K, Wu H, Pan F, Chen L, Feng C, Liu Y, et al. A machine learning-based model to predict acute traumatic coagulopathy in trauma patients upon emergency hospitalization. Clin Appl Thromb Hemost 2020;26:1076029619897827 [FREE Full text] [doi: 10.1177/1076029619897827] [Medline: 31908189]

54. Oh S, Park E, Jin Y, Piao J, Lee S. Automatic delirium prediction system in a Korean surgical intensive care unit. Nurs Crit Care 2014 Nov;19(6):281-291. [doi: 10.1111/nicc.12048] [Medline: 24165109]

55. Milbrandt EB, Clermont G, Martinez J, Kersten A, Rahim MT, Angus DC. Predicting late anemia in critical illness. Crit Care 2006 Feb;10(1):R39 [FREE Full text] [doi: 10.1186/cc4847] [Medline: 16507173]

56. Fialho AS, Celi LA, Cismondi F, Vieira SM, Reti SR, Sousa JMC, et al. Disease-based modeling to predict fluid response in intensive care units. Methods Inf Med 2013;52(6):494-502 [FREE Full text] [doi: 10.3414/ME12-01-0093] [Medline: 23986268]

57. Ghose S, Mitra J, Khanna S, Dowling J. An improved patient-specific mortality risk prediction in ICU in a random forest classification framework. Stud Health Technol Inform 2015;214:56-61. [Medline: 26210418]

58.   Venugopalan J, Chanani N, Maher K, Wang MD. Combination of static and temporal data analysis to predict mortality and readmission in the intensive care. Annu Int Conf IEEE Eng Med Biol Soc 2017 Jul;2017:2570-2573 [FREE Full text] [doi: 10.1109/EMBC.2017.8037382] [Medline: 29060424]

59.   Ting H, Chen M, Hsieh Y, Chan C. Good mortality prediction by Glasgow Coma Scale for neurosurgical patients. J Chin Med Assoc 2010 Mar;73(3):139-143 [FREE Full text] [doi: 10.1016/S1726-4901(10)70028-9] [Medline: 20230998]

60.   Sha Y, Wang MD. Interpretable predictions of clinical outcomes with an attention-based recurrent neural network. ACM BCB 2017 Aug;2017:233-240 [FREE Full text] [doi: 10.1145/3107411.3107445] [Medline: 32577628]

61.   Meiring C, Dixit A, Harris S, MacCallum NS, Brealey DA, Watkinson PJ, et al. Optimal intensive care outcome prediction over time using machine learning. PLoS One 2018;13(11):e0206862. [doi: 10.1371/journal.pone.0206862] [Medline: 30427913]

62.   Bukan RI, Møller AM, Henning MAS, Mortensen KB, Klausen TW, Waldau T. Preadmission quality of life can predict mortality in intensive care unit--a prospective cohort study. J Crit Care 2014 Dec;29(6):942-947. [doi: 10.1016/j.jcrc.2014.06.009] [Medline: 25060638]

63.   Hsieh Y, Su M, Wang C, Wang P. Prediction of survival of ICU patients using computational intelligence. Comput Biol Med 2014 Apr;47:13-19. [doi: 10.1016/j.compbiomed.2013.12.012] [Medline: 24508564]

64.   Oeyen S, Vermeulen K, Benoit D, Annemans L, Decruyenaere J. Development of a prediction model for long-term quality of life in critically ill patients. J Crit Care 2018 Feb;43:133-138. [doi: 10.1016/j.jcrc.2017.09.006] [Medline: 28892669]

65.   de Lange DW, Brinkman S, Flaatten H, Boumendil A, Morandi A, Andersen FH, VIP1 Study Group. Cumulative prognostic score predicting mortality in patients older than 80 years admitted to the ICU. J Am Geriatr Soc 2019 Jun;67(6):1263-1267 [FREE Full text] [doi: 10.1111/jgs.15888] [Medline: 30977911]

66.   Guidet B, de Lange DW, Boumendil A, Leaver S, Watson X, Boulanger C, VIP2 study group. The contribution of frailty, cognition, activity of daily life and comorbidities on outcome in acutely admitted patients over 80 years in European ICUs: the VIP2 study. Intensive Care Med 2020 Jan;46(1):57-69 [FREE Full text] [doi: 10.1007/s00134-019-05853-1] [Medline: 31784798]

67.   Heyland DK, Stelfox HT, Garland A, Cook D, Dodek P, Kutsogiannis J, Canadian Critical Care Trials Group and the Canadian Researchers at the End of Life Network. Predicting performance status 1 year after critical illness in patients 80 years or older: development of a multivariable clinical prediction model. Crit Care Med 2016 Sep;44(9):1718-1726. [doi: 10.1097/CCM.0000000000001762] [Medline: 27075141]

68.   Puskarich M. A decision tree incorporating biomarkers and patient characteristics estimates mortality risk for adults with septic shock. Evid Based Nurs 2015 Apr;18(2):42. [doi: 10.1136/eb-2014-101903] [Medline: 25163470]

69.   Wong HR, Lindsell CJ, Pettilä V, Meyer NJ, Thair SA, Karlsson S, et al. A multibiomarker-based outcome risk stratification model for adult septic shock*. Crit Care Med 2014 Apr;42(4):781-789 [FREE Full text] [doi: 10.1097/CCM.0000000000000106] [Medline: 24335447]

70.   Jaimes F, Farbiarz J, Alvarez D, Martínez C. Comparison between logistic regression and neural networks to predict death in patients with suspected sepsis in the emergency room. Crit Care 2005 Apr;9(2):R150-R156 [FREE Full text] [doi: 10.1186/cc3054] [Medline: 15774048]

71.   Ribas Ripoll VJ, Vellido A, Romero E, Ruiz-Rodríguez JC. Sepsis mortality prediction with the Quotient Basis Kernel. Artif Intell Med 2014 May;61(1):45-52. [doi: 10.1016/j.artmed.2014.03.004] [Medline: 24726036]

72.   Sha Y, Venugopalan J, Wang MD. A novel temporal similarity measure for patients based on irregularly measured data in electronic health records. ACM BCB 2016 Oct;2016:337-344 [FREE Full text] [doi: 10.1145/2975167.2975202] [Medline: 32577627]

73.   Yang T, Sun S, Zhao Y, Liu Q, Han M, Lin L, et al. Biomarkers upon discontinuation of renal replacement therapy predict 60-day survival and renal recovery in critically ill patients with acute kidney injury. Hemodial Int 2018 Jan;22(1):56-65. [doi: 10.1111/hdi.12532] [Medline: 28078828]

74.   Xu Z, Luo Y, Adekkanattu P, Ancker JS, Jiang G, Kiefer RC, et al. Stratified mortality prediction of patients with acute kidney injury in critical care. Stud Health Technol Inform 2019 Aug 21;264:462-466. [doi: 10.3233/SHTI190264] [Medline: 31437966]

75.   Trongtrakul K, Patumanond J, Kongsayreepong S, Morakul S, Pipanmekaporn T, Akaraborworn O, et al. Acute kidney injury risk prediction score for critically-ill surgical patients. BMC Anesthesiol 2020 Jun 03;20(1):140 [FREE Full text] [doi: 10.1186/s12871-020-01046-2] [Medline: 32493268]

76.   Bernal W, Wang Y, Maggs J, Willars C, Sizer E, Auzinger G, et al. Development and validation of a dynamic outcome prediction model for paracetamol-induced acute liver failure: a cohort study. Lancet Gastroenterol Hepatol 2016 Nov;1(3):217-225. [doi: 10.1016/S2468-1253(16)30007-3] [Medline: 28404094]

77.   Lindenmeyer CC, Flocco G, Sanghi V, Lopez R, Kim AJ, Niyazi F, et al. LIV-4: A novel model for predicting transplant-free survival in critically ill cirrhotics. World J Hepatol 2020 Jun 27;12(6):298-311 [FREE Full text] [doi: 10.4254/wjh.v12.i6.298] [Medline: 32742572]

78.   Balekian AA, Gould MK. Predicting in-hospital mortality among critically ill patients with end-stage liver disease. J Crit Care 2012 Dec;27(6):740.e1-740.e7 [FREE Full text] [doi: 10.1016/j.jcrc.2012.08.017] [Medline: 23059012]

79. Santos HGD, Zampieri FG, Normilio-Silva K, Silva GTD, Lima ACPD, Cavalcanti AB, et al. Machine learning to predict 30-day quality-adjusted survival in critically ill patients with cancer. J Crit Care 2020 Feb;55:73-78. [doi: 10.1016/j.jcrc.2019.10.015] [Medline: 31715534]

80. Vincent F, Soares M, Mokart D, Lemiale V, Bruneel F, Boubaya M, GrrrOH: Groupe de recherche respiratoire en réanimation en Onco-Hématologie (Group for respiratory research in intensive care in Onco-Hematology, http://www.grrroh.com/). In-hospital and day-120 survival of critically ill solid cancer patients after discharge of the intensive care units: results of a retrospective multicenter study-A Groupe de recherche respiratoire en réanimation en Onco-Hématologie (Grrr-OH) study. Ann Intensive Care 2018 Mar 27;8(1):40 [FREE Full text] [doi: 10.1186/s13613-018-0386-6] [Medline: 29582210]

81. Lee S, Zhao X, Davis KA, Topjian AA, Litt B, Abend NS. Quantitative EEG predicts outcomes in children after cardiac arrest. Neurology 2019 May 14;92(20):e2329-e2338 [FREE Full text] [doi: 10.1212/WNL.0000000000007504] [Medline: 30971485]

82. Murtuza B, Wall D, Reinhardt Z, Stickley J, Stumper O, Jones TJ, et al. The importance of blood lactate clearance as a predictor of early mortality following the modified Norwood procedure. Eur J Cardiothorac Surg 2011 Nov;40(5):1207-1214. [doi: 10.1016/j.ejcts.2011.01.081] [Medline: 21450476]

83. Gracia Arnillas MP, Alvarez-Lerma F, Masclans J, Roquer J, Soriano C, Manzano D, et al. Impact of adrenomedullin levels on clinical risk stratification and outcome in subarachnoid haemorrhage. Eur J Clin Invest 2020 Nov;50(11):e13318. [doi: 10.1111/eci.13318] [Medline: 32535893]

84. Haveman ME, Van Putten MJAM, Hom HW, Eertman-Meyer CJ, Beishuizen A, Tjepkema-Cloostermans MC. Predicting outcome in patients with moderate to severe traumatic brain injury using electroencephalography. Crit Care 2019 Dec 11;23(1):401 [FREE Full text] [doi: 10.1186/s13054-019-2656-6] [Medline: 31829226]

85. Wildman MJ, Sanderson C, Groves J, Reeves BC, Ayres J, Harrison D, et al. Predicting mortality for patients with exacerbations of COPD and asthma in the COPD and asthma outcome study (CAOS). QJM 2009 Jun;102(6):389-399. [doi: 10.1093/qjmed/hcp036] [Medline: 19369483]

86. Daly K, Beale R, Chang RW. Reduction in mortality after inappropriate early discharge from intensive care unit: logistic regression triage model. BMJ 2001 May 26;322(7297):1274-1276 [FREE Full text] [doi: 10.1136/bmj.322.7297.1274] [Medline: 11375229]

87. Hernández-Tejedor A, Cabré-Pericas L, Martín-Delgado MC, Leal-Micharet AM, Algora-Weber A, EPIPUSE study group. Evolution and prognosis of long intensive care unit stay patients suffering a deterioration: a multicenter study. J Crit Care 2015 Jun;30(3):654.e1-654.e7. [doi: 10.1016/j.jcrc.2015.01.011] [Medline: 25656920]

88. Ji S, Smith R, Huynh T, Najarian K. A comparative analysis of multi-level computer-assisted decision making systems for traumatic injuries. BMC Med Inform Decis Mak 2009 Jan 14;9:2 [FREE Full text] [doi: 10.1186/1472-6947-9-2] [Medline: 19144188]

89. Che Z, Purushotham S, Khemani R, Liu Y. Interpretable deep models for ICU outcome prediction. AMIA Annu Symp Proc 2016;2016:371-380 [FREE Full text] [Medline: 28269832]

90. Ebadollahi S, Sun J, Gotz D, Hu J, Sow D, Neti C. Predicting patient's trajectory of physiological data using temporal trends in similar patients: a system for near-term prognostics. AMIA Annu Symp Proc 2010 Nov 13;2010:192-196 [FREE Full text] [Medline: 21346967]

91. Castiñeira D, Schlosser KR, Geva A, Rahmani AR, Fiore G, Walsh BK, et al. Adding continuous vital sign information to static clinical data improves the prediction of length of stay after intubation: a data-driven machine learning approach. Respir Care 2020 Sep;65(9):1367-1377. [doi: 10.4187/respcare.07561] [Medline: 32879034]

92. Mueller M, Wagner CL, Annibale DJ, Knapp RG, Hulsey TC, Almeida JS. Parameter selection for and implementation of a web-based decision-support tool to predict extubation outcome in premature infants. BMC Med Inform Decis Mak 2006 Mar 01;6:11 [FREE Full text] [doi: 10.1186/1472-6947-6-11] [Medline: 16509967]

93. Mueller M, Wagner CL, Annibale DJ, Hulsey TC, Knapp RG, Almeida JS. Predicting extubation outcome in preterm newborns: a comparison of neural networks with clinical expertise and statistical modeling. Pediatr Res 2004 Jul;56(1):11-18. [doi: 10.1203/01.PDR.0000129658.55746.3C] [Medline: 15128922]

94. Dunning J, Au J, Kalkat M, Levine A. A validated rule for predicting patients who require prolonged ventilation post cardiac surgery. Eur J Cardiothorac Surg 2003 Aug;24(2):270-276. [doi: 10.1016/s1010-7940(03)00269-0] [Medline: 12895619]

95. Manji RA, Arora RC, Singal RK, Hiebert B, Moon MC, Freed DH, et al. Long-term outcome and predictors of noninstitutionalized survival subsequent to prolonged intensive care unit stay after cardiac surgical procedures. Ann Thorac Surg 2016 Jan;101(1):56-63; discussion 63. [doi: 10.1016/j.athoracsur.2015.07.004] [Medline: 26431924]

96. Brandi S, Troster EJ, Cunha MLDR. Length of stay in pediatric intensive care unit: prediction model. Einstein (Sao Paulo) 2020;18:eAO5476 [FREE Full text] [doi: 10.31744/einstein_journal/2020AO5476] [Medline: 33053018]

97. McWilliams CJ, Lawson DJ, Santos-Rodriguez R, Gilchrist ID, Champneys A, Gould TH, et al. Towards a decision support tool for intensive care discharge: machine learning algorithm development using electronic healthcare data from MIMIC-III and Bristol, UK. BMJ Open 2019 Mar 07;9(3):e025925 [FREE Full text] [doi: 10.1136/bmjopen-2018-025925] [Medline: 30850412]

98. Lin Y, Zhou Y, Faghri F, Shaw MJ, Campbell RH. Analysis and prediction of unplanned intensive care unit readmission using recurrent neural networks with long short-term memory. PLoS One 2019;14(7):e0218942 [FREE Full text] [doi: 10.1371/journal.pone.0218942] [Medline: 31283759]

99. Czeiter E, Amrein K, Gravesteijn BY, Lecky F, Menon DK, Mondello S, CENTER-TBI Participants and Investigators. Blood biomarkers on admission in acute traumatic brain injury: relations to severity, CT findings and care path in the CENTER-TBI study. EBioMedicine 2020 Jun;56:102785 [FREE Full text] [doi: 10.1016/j.ebiom.2020.102785] [Medline: 32464528]

100. Yin W, Li Y, Zeng X, Qin Y, Wang D, Zou T, et al. The utilization of critical care ultrasound to assess hemodynamics and lung pathology on ICU admission and the potential for predicting outcome. PLoS One 2017;12(8):e0182881 [FREE Full text] [doi: 10.1371/journal.pone.0182881] [Medline: 28806783]

101. Shickel B, Loftus TJ, Adhikari L, Ozrazgat-Baslanti T, Bihorac A, Rashidi P. DeepSOFA: a continuous acuity score for critically ill patients using clinically interpretable deep learning. Sci Rep 2019 Feb 12;9(1):1879 [FREE Full text] [doi: 10.1038/s41598-019-38491-0] [Medline: 30755689]

102. McRae MP, Simmons GW, Christodoulides NJ, Lu Z, Kang SK, Fenyo D, et al. Clinical decision support tool and rapid point-of-care platform for determining disease severity in patients with COVID-19. Lab Chip 2020 Jun 21;20(12):2075-2085. [doi: 10.1039/d0lc00373e] [Medline: 32490853]

103. Sanchez-Pinto LN, Luo Y, Churpek MM. Big data and data science in critical care. Chest 2018 Nov;154(5):1239-1248. [doi: 10.1016/j.chest.2018.04.037]

104. Liu Y, Chen PHC, Krause J, Peng L. How to read articles that use machine learning: users' guides to the medical literature. JAMA 2019 Nov 12;322(18):1806-1816. [doi: 10.1001/jama.2019.16489] [Medline: 31714992]

105. Riley RD, Ensor J, Snell KIE, Harrell FE, Martin GP, Reitsma JB, et al. Calculating the sample size required for developing a clinical prediction model. BMJ 2020 Mar 18;368:m441. [doi: 10.1136/bmj.m441] [Medline: 32188600]

106. Collins GS, Reitsma JB, Altman DG, Moons KGM. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): the TRIPOD statement. Br J Surg 2015 Feb;102(3):148-158. [doi: 10.1002/bjs.9736] [Medline: 25627261]

107. Ho SY, Phua K, Wong L, Bin Goh WW. Extensions of the external validation for checking learned model interpretability and generalizability. Patterns 2020 Nov;1(8):100129. [doi: 10.1016/j.patter.2020.100129]

108. Futoma J, Simons M, Panch T, Doshi-Velez F, Celi LA. The myth of generalisability in clinical research and machine learning in health care. Lancet Digit Health 2020 Sep;2(9):e489-e492 [FREE Full text] [doi: 10.1016/S2589-7500(20)30186-2] [Medline: 32864600]

Abbreviations

EEG: electroencephalography

Review

Machine Learning–Based Short-Term Mortality Prediction Models for Patients With Cancer Using Electronic Health Record Data: Systematic Review and Critical Appraisal

Sheng-Chieh Lu[1], PhD; Cai Xu[1], PhD; Chandler H Nguyen[2], BS; Yimin Geng[3], MSLS; André Pfob[4], MD; Chris Sidey-Gibbons[1], PhD

[1]Department of Symptom Research, The University of Texas MD Anderson Cancer Center, Houston, TX, United States
[2]McGovern Medical School, University of Texas Health Science Center, Houston, TX, United States
[3]Research Medical Library, The University of Texas MD Anderson Cancer Center, Houston, TX, United States
[4]Department of Obstetrics and Gynecology, Heidelberg University Hospital, Heidelberg, Germany

Corresponding Author:
Chris Sidey-Gibbons, PhD
Department of Symptom Research
The University of Texas MD Anderson Cancer Center
6565 MD Anderson Boulevard
Houston, TX, 77030
United States
Phone: 1 713 794 4453
Fax: 1 713 745 3475
Email: cgibbons@mdanderson.org

Abstract

Background: In the United States, national guidelines suggest that aggressive cancer care should be avoided in the final months of life. However, guideline compliance currently requires clinicians to make judgments based on their experience as to when a patient is nearing the end of their life. Machine learning (ML) algorithms may facilitate improved end-of-life care provision for patients with cancer by identifying patients at risk of short-term mortality.

Objective: This study aims to summarize the evidence for applying ML in ≤1-year cancer mortality prediction to assist with the transition to end-of-life care for patients with cancer.

Methods: We searched MEDLINE, Embase, Scopus, Web of Science, and IEEE to identify relevant articles. We included studies describing ML algorithms predicting ≤1-year mortality in patients of oncology. We used the prediction model risk of bias assessment tool to assess the quality of the included studies.

Results: We included 15 articles involving 110,058 patients in the final synthesis. Of the 15 studies, 12 (80%) had a high or unclear risk of bias. The model performance was good: the area under the receiver operating characteristic curve ranged from 0.72 to 0.92. We identified common issues leading to biased models, including using a single performance metric, incomplete reporting of or inappropriate modeling practice, and small sample size.

Conclusions: We found encouraging signs of ML performance in predicting short-term cancer mortality. Nevertheless, no included ML algorithms are suitable for clinical practice at the current stage because of the high risk of bias and uncertainty regarding real-world performance. Further research is needed to develop ML models using the modern standards of algorithm development and reporting.

XSL•FO
RenderX

Introduction

Background

Cancer therapies, including chemotherapy, immunotherapy, radiation, and surgery, aim to cure and reduce the risk of recurrence in early-stage disease and improve survival and quality of life for late-stage disease. However, cancer therapy is invariably associated with negative effects, including toxicity, comorbidities, financial burden, and social disruption. There is growing recognition that therapies are sometimes started too late, and many patients die while receiving active therapy [1-3]. For instance, a systematic review summarized that the percentage of patients with lung cancer receiving aggressive treatments during the last month of their life ranged from 6.4% to >50% [4]. Another retrospective comparison study revealed that the proportion of patients with gynecologic cancer undergoing chemotherapy or invasive procedures in their last 3 months was significantly higher in 2011 to 2015 than in 2006 to 2010 [5]. Research has shown that the aggressiveness of care at the end of life in patients with advanced cancers is associated with extra costs and a reduction in the quality of life for patients and their families [4,6].

In the United States, national guidelines state that gold standard cancer care should avoid the provision of aggressive care in the final months of life [7]. Avoiding aggressive care at the end of life currently requires clinicians to make judgments based on their experience as to when a patient is nearing the end of their life [8]. Research has shown that these decisions are difficult to make because of a lack of scientific, objective evidence to support the clinicians' judgment in palliative or related discussion initiation [2,9,10]. Thus, a decision support tool enabling the early identification of patients of oncology who may not benefit from aggressive care is needed to support better palliative care management and reduce clinicians' burden [2].

In recent years, there have been substantial changes in both the type and quantity of patient data collected using electronic health records (EHR) and the sophistication and availability of the techniques used to learn the complex patterns within that data. By learning these patterns, it is possible to make predictions for individual patients' future health states [11]. The process of creating accurate predictions from evident patterns in past data is referred to as machine learning (ML), a branch of artificial intelligence research [12]. There has been growing enthusiasm for the development of ML algorithms to guide clinical problems. Using ML to create robust, individualized predictions of clinical outcomes, such as the risk of short-term mortality [13,14], may improve care by allowing clinical teams to adjust care plans in anticipation of a forecasted event. Such predictions have been shown to be acceptable for use in clinical practice [15] and may one day become a fundamental aspect of clinical practice.

ML applications have been developed to support mortality predictions for a variety of populations, including but not limited to patients with traumatic brain injury, COVID-19 disease of 2019, and cancers, as well as patients admitted to emergency departments and intensive care units. These applications have consistently demonstrated promising performances across studies [16-19]. Researchers have also applied ML techniques to create tools supporting various clinical tasks involved in the care of patients of oncology, with most applications focusing on the prediction of cancer susceptibility, recurrence, treatment response, and survival [14,19,20]. However, the performance of ML applications in supporting mortality predictions for patients of oncology has not yet been systematically examined and synthesized.

In addition, as the popularity of ML in clinical medicine has risen, so too has the realization that applying complex algorithms to big data sets does not in itself result in high-quality models [11,21]. For example, subtle temporal-regional nuances in data can cause models to learn relationships that are not repeated over time and space. This can lead to poor future performance and misleading predictions [22]. Algorithms may also learn to replicate human biases in data and, as a result, could produce predictions that negatively affect disadvantaged groups [23,24]. Recent commentary has drawn attention to various issues in the transparency, performance, and reproducibility of ML tools [25-27]. A comparison of 511 scientific papers describing the development of ML algorithms found that, in terms of reproducibility, ML for health care compared poorly to other fields [28]. Issues of algorithmic fairness and performance are especially pertinent when predicting patient mortality. If done correctly, these predictions could help patients and their families receive gold standard care at the end of life; if done incorrectly, there is a risk of causing unnecessary harm and distress at a deeply sensitive time.

Another aspect of mortality affecting the algorithm performance is its rare occurrence in most populations. There are known issues that are commonly encountered when trying to predict events from data sets in which there are far fewer events than nonevents, which is known as class imbalance. One such issue is known as the *accuracy paradox*—the case in which an ML algorithm presents with high accuracy but a failure to identify occurrences of the rare outcome it was tasked to predict [29,30]. During the model training process, many algorithms seek to maximize their accuracy across the entire data set. In the case of a data set in which only 10% of patients experienced a rare outcome—as is often the case with data sets containing mortality—an algorithm could achieve an apparently excellent accuracy of 0.90 by simply predicting that every patient would live. The resulting algorithm would be clinically useless on account of its failure to identify patients who are at risk of dying. If handled incorrectly, the class imbalance problem can lead algorithms to prioritize the predictions of the majority class. For this reason, it is especially important to evaluate multiple performance metrics when assessing algorithms that predict rare events.

Objective

The purpose of this systematic review is to critically evaluate the current evidence to (1) summarize ML-based model performance in predicting ≤1-year mortality for patients with cancer, (2) evaluate the practice and reporting of ML modeling, and (3) provide suggestions to guide future work in the area. In this study, we seek to evaluate models identifying patients with cancer who are near the end of their life and may benefit from

end-of-life care to facilitate the better provision of care. As the definitions of aggressive care at the end of life vary from initiation of chemotherapy or invasive procedures or admission to the emergency department or intensive care unit within 14 days to 6 months [1,4,5], we focused on ≤1-year mortality of patients with cancer to ensure that we include all ML models that have the potential to reduce the aggressiveness of care and support the better provision of palliative care for cancer populations.

Methods

Overview

We conducted this systematic review following the Joanna Briggs Institute guidelines for systematic reviews [31]. To facilitate reproducible reporting, we present our results following the PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) statement [32]. This review was prospectively registered in PROSPERO (International Prospective Register of Systematic Reviews; PROSPERO ID: CRD42021246233). The protocol for this review has not been published.

Search Strategy

We searched Ovid MEDLINE, Ovid Embase, Clarivate Analytics Web of Science, Elsevier Scopus, and IEEE Xplore databases from the date of inception to October 2020. The following concepts were searched using subject headings keywords as needed: *cancer*, *tumor*, *oncology*, *machine learning*, *artificial intelligence*, *performancemetrics*, *mortality*, *cancer death*, *survival rate*, and *prognosis*. The terms were combined using AND/OR Boolean statements. A full list of search terms along with a complete search strategy for each database used is provided in Multimedia Appendix 1. In addition, we reviewed the reference lists of each included study for relevant studies.

Study Selection

A total of 2 team members screened all the titles and abstracts of the articles identified in the search for studies. A senior ML researcher (CSG) resolved the discrepancies between the 2 reviewers. We then examined the full text of the remaining articles using the same approach but resolved disagreements via consensus. Studies were included if they (1) developed or validated ML-based models predicting ≤1-year mortality for patients of oncology, (2) made predictions using EHR data, (3) reported model performance, and (4) were original research published through a peer-reviewed process in English. We excluded studies if they (1) focused on risk factor investigation; (2) implemented existing models; (3) were not specific to patients with cancer; (4) used only image, genomic, clinical trial, or publicly available data; (5) predicted long-term (>1 year) mortality or survival probability; (6) created survival stratification using unsupervised ML approaches; and (7) were not peer-reviewed full papers. We defined short-term mortality as death happening within ≤1 year after receiving cancer diagnostics or certain treatments for this review.

Critical Appraisal

We evaluated the risk of bias (ROB) of each included study using the prediction model ROB assessment tool [33]. A total of 2 reviewers independently conducted the assessment for all the included studies and resolved conflicts by consensus.

Data Extraction and Synthesis

For data extraction, we developed a spreadsheet based on the items in the transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD) [34] through iterative discussions. A total of 4 reviewers independently extracted information about sampling, data sources, predictive and outcome variables, modeling and evaluation approaches, model performance, and model interpretations using the spreadsheet from the included studies, with each study extracted by 2 reviewers. Discrepancies were discussed among all reviewers to reach a consensus. The collected data items are available in Multimedia Appendix 2 [35]. To summarize the evidence, we grouped the studies using TRIPOD's classification for prediction model studies (Textbox 1) and summarized the data narratively and descriptively by group. To estimate the performance of each ML algorithm, we averaged the area under the receiver operating characteristic curve (AUROC) for each type of ML algorithm across the included studies and estimated SE for 95% CI calculation using the averaged AUROC and pooled validation sample size for each type of ML algorithm. In addition, we conducted a sensitivity analysis to assess the impact of studies that were outliers either on the basis of their sample size or their risk of bias.

Textbox 1. Types of prediction model studies.

Study type and definition

Type 1a

Studies develop prediction model or models and evaluate model performance using the same data used for model development.

Type 1b

Studies develop prediction model or models and evaluate the model or models using the same data used for model development with resampling techniques (eg, bootstrapping and cross-validation) to avoid an optimistic performance estimate.

Type 2a

Studies randomly split data into two subsets: one for model development and another for model performance estimate.

Type 2b

Studies nonrandomly split data into two subsets: one for model development and another for model performance estimate. The splitting rule can be by institute, location, and time.

Type 3

Studies develop and evaluate prediction model or models using 2 different data sets (eg, from different studies).

Type 4

Studies evaluate existing prediction models with new data sets not used in model development.

Note: The types of prediction model studies were summarized from Collins et al [34].

Results

Summary of Included Studies

Our search resulted in 970 unduplicated references, of which we excluded 771 (79.5%) articles because of various reasons, such as no ML involvement, not using EHR data, or no patient with cancer involvement, based on the title and abstract screen. After the full-text review, we included 1.5% (15/970) of articles involving a total of 110,058 patients with cancer (Figure 1). We have provided a detailed record of the selection process in Multimedia Appendix 3 [36,37].

Figure 1. PRISMA (Preferred Reporting Item for Systematic Reviews and Meta-Analyses) flowchart diagram for the study selection process. ML: machine learning.
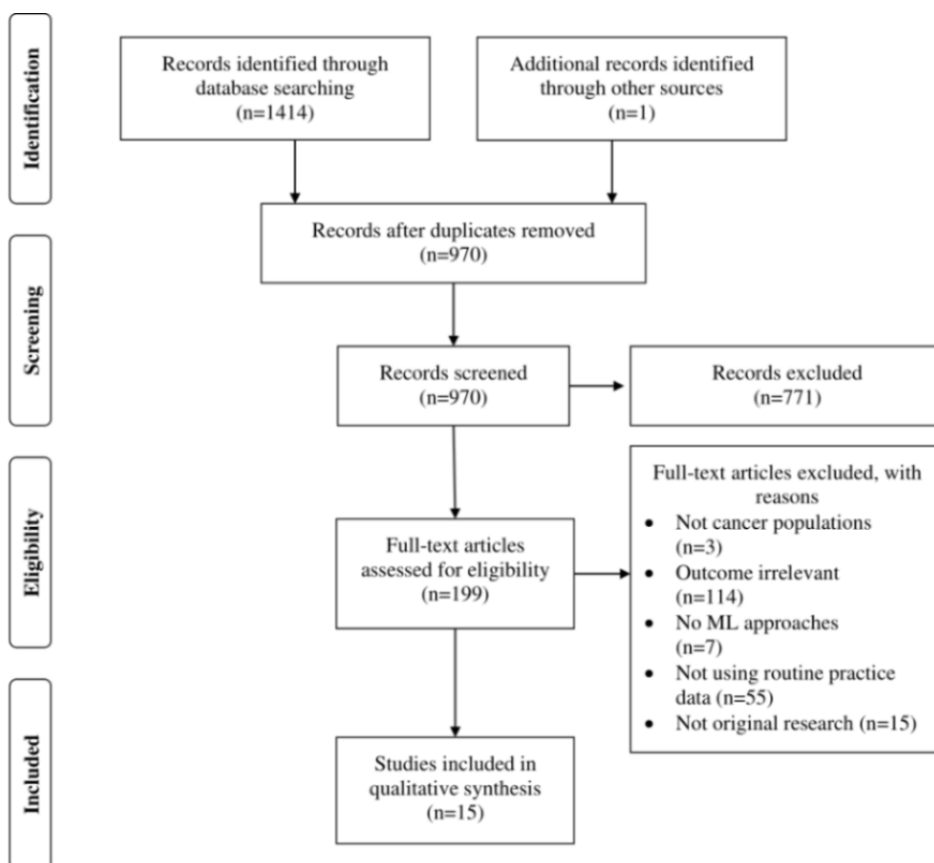
We present a characteristic summary of the included articles in Table 1 [36-49]. Of the 15 included articles, 13 (87%) were model development and internal validations, and 2 (13%) were external validations of existing models. The median sample size was 783 (range 173-26,946), with a median of 21 predictors considered (range 9-5390). The target populations of the 15 articles included 5 (33%) with all types of cancer, 3 (20%) with spinal metastatic diseases, 2 (13%) with liver cancer, and 1 (7%) each with gastric cancer, colon and rectum cancer, stomach cancer, lung cancer, and bladder cancer. Several algorithms have been examined in many studies. The most commonly used ML algorithms were artificial neural networks (8/15, 53%). Other algorithms included gradient-boosted trees (4/15, 27%), decision trees (4/15, 27%), regularized logistic regression (LR; 4/15, 27%), stochastic gradient boosting (2/15, 13%), naive Bayes classifier (1/15, 7%), Bayes point machine (1/15, 7%), and random forest (RF; 1/15, 7%). Of the 15 studies, 2 (13%) tested their models in their training data sets by resampling (type 1b), 9 (60%) examined their models using randomly split holdout internal validation data sets (type 2a), 2 (13%) examined with nonrandomly split holdout validation data sets (type 2b), and 2 (13%) validated existing models using external data sets (type 4). The frequent candidate predictors were demographic (12/15, 80%), clinicopathologic (12/15, 80%), tumor entity (7/15, 47%), laboratory (7/15, 47%), comorbidity (5/15, 33%), and prior treatment information (5/15, 33%). The event of interest varied across the studies, with 47% (7/15) for 1-year mortality, 33% (5/15) for 180-day mortality, 13% (2/15) for 90-day mortality, and 7% (1/15) for 30-day mortality.

Table 1. Characteristics of the included studies (N=15).

Type of cancer and study	Country	Study type	Treatment	Sample size			Algorithms	Input features (total number of features)	Outcome
				Training	Testing	Validating			
All cancer									
Sena et al [38]	Brazil	1b	All	543	N/A[a]	N/A	DT[b], ANN[c], and NB[d]	Comorbidity and PRO[e] for physical and mental status assessments (9)	180-day death
Parikh et al [39]	United States	2a	All	18,567	7958	N/A	GBT[f] and RF[g]	Demographic, clinico-pathologic, laboratory, comorbidity, and electrocardiogram data (599)	180-day death
Manz et al [37]	United States	4	All	N/A	N/A	24,582	GBT	Same as Parikh et al [39]	180-day death
Bertsimas et al [50]	United States	2a	All	14,427	9556	N/A	DT, regularized LR[h], and GBT	Demographic, clinico-pathologic, gene mutations, prior treatment, comorbidity, use of health care resources, vital signs, and laboratory data (401)	180-day death
Elfiky et al [43]	United States	2b	All	17,832	9114	N/A	GBT	Demographic, clinico-pathologic, prescription, comorbidity, laboratory, vital sign, and use of health care resources data and physician notes (5390)	180-day death
Non–small cell lung cancer									
Hanai et al [44]	Japan	2b	Curative resection	125	48	N/A	ANN	Demographic, clinico-pathologic, and tumor entity data (17)	1-year death
Gastric cancer									
Nilsaz-Dez-fouli et al [45]	Iran	1b	Surgery	452	N/A	N/A	ANN	Demographic, clinico-pathologic, tumor entity, and prior treatment (20)	1-year death
Colon and rectum cancer									
Arostegui et al [46]	Spain	2a	Curative or palliative surgery	981	964	N/A	DT and regularized LR	Demographic, clinico-pathologic, tumor entity, comorbidity, ASA[i] prior treatment, laboratory, operational data, postoperational complication, and use of health care resources data (32)	1-year death
Stomach cancer									
Biglarian et al [47]	Iran	2a	Surgery	300	136	N/A	ANN	Demographic, clinico-pathologic, and symptom data (NR[j])	1-year death
Bladder cancer									

XSL•FO
RenderX

Type of cancer and study	Country	Study type	Treatment	Sample size			Algorithms	Input features (total number of features)	Outcome
				Training	Testing	Validating			
Klén et al [48]	Turkey	2a	Radical cystectomy	733	366	N/A	Regularized LR	Demographic, clinico-pathologic, ASA, co-morbidity, laboratory, prior treatment, tomography, and operational data (NR)	90-day death
Hepatocellular carcinoma									
Chiu et al [49]	Taiwan	2a	Liver resection	347	87	N/A	ANN	Demographic, clinico-pathologic, tumor entity, comorbidity, ASA, laboratory, opera-tional, and postopera-tional data (21)	1-year death
Zhang et al [40]	China	2a	Liver transplant	230	60	N/A	ANN	Donor demographic data and recipient lab-oratory, clinicopatho-logic, and image data (14)	1-year death
Spinal metastatic									
Karhade et al [41]	United States	2a	Surgery	1432	358	N/A	ANN, SVM[k], DT, and BPM[l]	Demographic, clinico-pathologic, tumor enti-ty, ASA, laboratory, and operational data (23)	30-day death
Karhade et al [42]	United States	2a	Surgery	587	145	N/A	SGB[m], RF, ANN, SVM, and regular-ized LR	Demographic, clinico-pathologic, tumor enti-ty, laboratory, opera-tional, ECOG[n], ASIA[o], and prior treatment data (29)	90-day death
Karhade et al [36]	United States	4	Curative surgery	N/A	N/A	176	SGB	ECOG, demographic, clinicopathologic, tu-mor entity, laboratory, prior treatment, and ASIA data (23)	1-year death

[a]N/A: not applicable.

[b]DT: decision tree.

[c]ANN: artificial neural network.

[d]NB: naive Bayes.

[e]PRO: patient-reported outcome.

[f]GBT: gradient-boosted tree.

[g]RF: random forest.

[h]LR: logistic regression.

[i]ASA: American Sociological Association.

[j]NR: not reported.

[k]SVM: support vector machine.

[l]BPM: Bayes point machine.

[m]SGB: stochastic gradient boosting.

[n]ECOG: Eastern Cooperative Oncology Group.

[o]ASIA: American Spinal Injury Association.

ROB Evaluation

Of the 15 studies, 12 (80%) were deemed to have a high or unclear ROB. The analysis domain was the major source of bias (Figure 2). Of the 12 model development studies, 8 (67%) provided insufficient or no information on data preprocessing and model optimization (tuning) methods. Approximately 33% (5/15) of studies did not report how they addressed missing

data, and 13% (2/15) potentially introduced selection bias by excluding patients with missing data. All studies clearly defined their study populations and data sources, although none justified their sample size. Predictors and outcomes of interest were also well-defined in all studies, except for 20% (3/15) of studies that did not specify their outcome measure definition and whether the definition was consistently used.

Figure 2. Risk of bias assessment for the included studies. Risk of bias assessment result for each included study using prediction model risk of bias assessment tool [15,35-49].

Model Performance

We summarize the performance of the best models from the type 2, 3, and 4 studies (12/15,80%) in Table 2. We excluded 1 type 2b study as the authors did not report their performance results in a holdout validation set. Model performance across the studies ranged from acceptable to good, based on AUROC ranging from 0.72 to 0.92. Approximately 40% (6/15) of studies reported only the AUROC values, therefore, leaving some uncertainty about model performance in correctly identifying patients at risk of short-term mortality. Other performance metrics were less reported and were sometimes indicative of poor performance. Studies reported median accuracy 0.91 (range 0.86-0.96; 2/15, 13%), sensitivity 0.85 (range 0.27-0.91; 4/15, 27%), specificity 0.90 (0.50-0.99; 5/15, 33%), as well as the positive predictive value (PPV) and the negative predictive value of 0.52 (range 0.45-0.83; 4/15, 27%) and 0.92 (range 0.86-0.97; 2/15, 13%), respectively.

Among the ML algorithms examined, all algorithms were similarly performed, with RF slightly better than the other algorithms (Figure 3). Approximately 33% (5/15) of studies compared their ML algorithms with statistical models [39,47,48,50,51]. Differences in AUROC between the ML and statistical models ranged from 0.01 to 0.11, with one of the studies reporting a significant difference (Table 2).

Table 2. Predicting performance for the best model for each study in a holdout internal or external validation data set (N=12).

Type of cancer and study	Outcome	Training sample	Validation sample	Mortality rate (%)	Algorithm	AUROC[a]	Accuracy	Sensitivity	Specificity	PPV[b]	NPV[c]	Calibration	Benchmark, model (Δ AUROC)
All cancer													
Manz et al [37]	180-day death	N/A[d]	24,582	4.2	GBT[e]	0.89	__[f]	0.27	0.99	0.45	0.97	Well-fit	—
Parikh et al [39]	180-day death	18,567	7958	4.0	RF[g]	0.87	0.96	—	0.99	0.51	—	Well-fit at the low-risk group	LR[h] (0.01)
Bertsimas et al [50]	180-day death	14,427	9556	5.6	GBT	0.87	0.87	.60	—	0.53	—	—	LR (0.11)
Elfiky et al [43]	180-day death	17,832	9114	18.4	GBT	0.83	—	—	—	—	—	Well-fit	—
Gastrointestinal cancer													
Arostegui et al [46]	1-year death	981	964	5.1	DT[i]	0.84	—	—	—	—	—	Well-fit	—
Biglarian et al [47]	1-year death	300	136	37.5	ANN[j]	0.92	—	0.80	0.85	—	—	—	CPH[k] (0.04)[l]
Patients with bladder cancer													
Klén et al [48]	90-day death	733	366	4.4	Regularized LR	0.72	—	—	—	—	—	—	ACCI[m] univariate model (0.05)
Patients with liver cancer													
Chiu et al [49]	1-year death	347	87	17	ANN	0.88	—	0.89	0.50	—	—	—	LR (0.08)
Zhang et al [40]	1-year death	230	60	23.9	ANN	0.91	—	0.91	0.90	0.83	0.86	—	—
Patients with spinal metastasis													
Karhade et al [41]	30-day death	1432	358	8.5	BPM[n]	0.78	—	—	—	—	—	Well-fit	—
Karhade et al [42]	1-year death	586	145	54.3	SGB[o]	0.89	—	—	—	—	—	Well-fit	—
Karhade et al [36]	1-year death	N/A	176	56.2	SGB	0.77	—	—	—	—	—	Fairly well-fit	—

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

[b]PPV: positive predictive value.

[c]NPV: negative predictive value.

[d]N/A: not applicable.

[e]GBT: gradient-boosted tree.

[f]No data available

[g]RF: random forest.

[h]LR: logistic regression.

[i]DT: decision tree.

[j]ANN: artificial neural network.

[k]CPH: Cox proportional hazard.

[l]Significant at the α level defined by the study.

[m]ACCI: adjusted Charlson comorbidity index.

[n]BPM: Bayes point machine

[o]SGB: stochastic gradient boosting.

XSL•FO
RenderX

Figure 3. Pooled AUROC by machine learning (ML) algorithm. ANN: artificial neural network; AUROC: area under the receiver operating characteristic curve; BPM: Bayes point machine; DT: decision tree; GBT: gradient-boosted tree; LR: logistic regression; RF: random forest; SGB: stochastic gradient boosting; SVM: support vector machine.
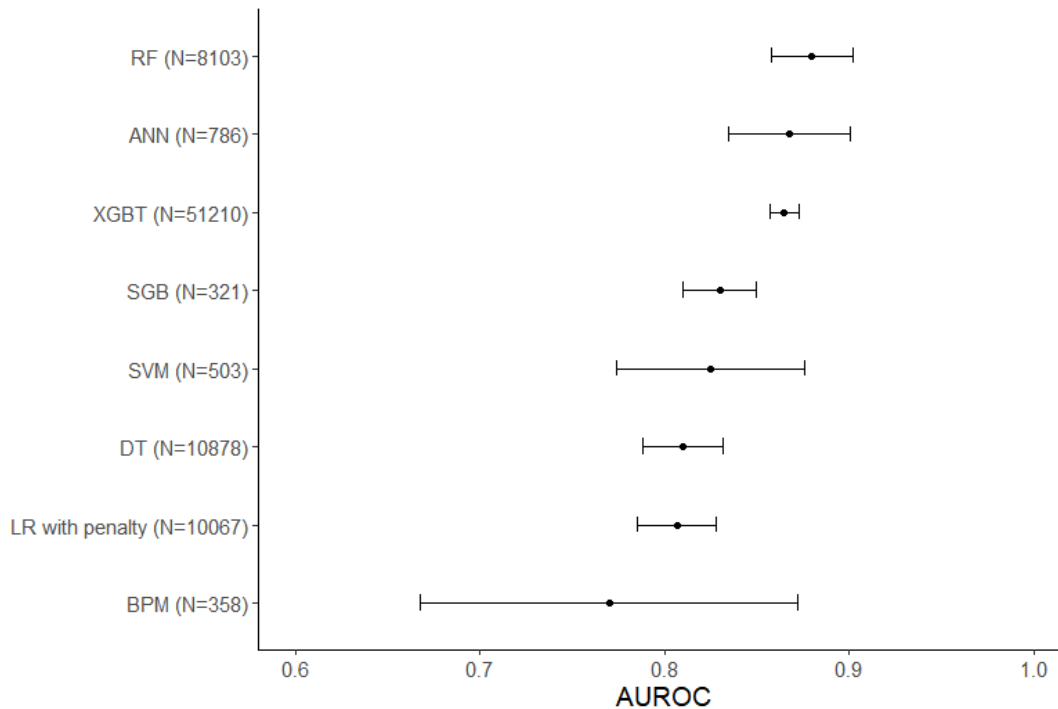
Model Development and Evaluation Processes

Most articles (11/15, 73%) did not report how their training data were preprocessed (Table 3). Authors of 27% (4/15) of articles reported their methods for preparing numeric variables, with 75% (3/4) using normalization, 25% (1/4) using standardization, and 25% (1/4) using discretization. Approximately 13% (2/15) of articles used one-hot encoding for their categorical variables. Various techniques were used to address missing data, including constant value imputation (3/15, 20%), multiple imputation (3/15, 20%), complete cases only (2/15, 13%), probabilistic imputation (1/15, 7%), and the optimal impute algorithm (1/15, 7%).

Of the 13 model development studies, 9 (69%) reported their approaches for feature selection. The approaches, including 3 model-based variable importance, between-variable correlation, zero variance, univariate Cox proportional hazard, forward stepwise selection algorithm, recursive feature selection, and parameter-increasing method, were used alone or in combination. Concerning hyperparameter selection, 33% (5/15) reported their methods to determine hyperparameters, with 60% (3/5) using grid search and 2 (40%) using the default values of the modeling software. Finally, 47% (7/15) used various resampling approaches to ensure the generalizability of their models. The N-fold cross-validation approach was the primary strategy. Varying fold numbers were used, such as 10 (3/15, 20%), 5 (2/15, 13%), 4 (1/15, 7%), 3 repeats 10 (1/15, 7%), and 5 repeats 5-fold (1/15, 7%). One of the studies used the bootstrapping method. Approximately 27% (4/15) of studies did not report whether resampling was performed.

Of the 15 studies, 12 (80%) used variable importance plots to interpret their models, 3 (20%) included decision tree rules, and 2 (13%) included coefficients to explain their models in terms of prediction generation. Other model interpretation approaches, including local interpretable model-agnostic explanations and partial dependence plots, were used in 7% (1/15) of studies.

Table 3. The Model development processes and evaluations used in the included studies.

Type and study	Data preprocessing			Model optimization			Interpretation
	Numeric variables	Categorical variables	Missing data	Feature selection	Hyperparameter value selection	Generalizability consideration	
Type 1b							
Sena et al [38]	Normalization	N/A[a]	NR[b]	None	Software default	10-fold CV[c]	VI[d]
Nilsaz-Dez-fouli et al [45]	NR	NR	NR	VI	Grid search	5×5-fold CV	VI
Type 2a							
Parikh et al [39]	NR	NR	Constant value imputation	Zero variance and between-variable correlation	Grid search	5-fold CV	VI and coefficient
Klén et al [48]	NR	NR	Complete cases only	LASSO[e] LR[f]	NR	NR	VI
Karhade et al [42]	NR	NR	missForest multiple imputation	RF[g]	NR	3×10-fold CV	VI, PDP[h], and LIME[i]
Karhade et al [41]	NR	NR	Multiple imputation	Recursive feature selection	NR	10-fold CV	NR
Arostegui et al [46]	Discretization	One-hot encoding	Constant value imputation	RF variable importance	Software default	Bootstrapping	VI and decision tree rules
Bertsimas et al [50]	NR	NR	Optimal impute algorithm	None	NR	NR	VI and decision tree rules
Chiu et al [49]	NR	NR	Complete cases only	Univariate Cox proportional hazard model	NR	NR	VI
Zhang et al [40]	Normalization	One-hot encoding	NR	Forward stepwise selection algorithm	NR	10-fold CV	VI
Biglarian et al [47]	NR	NR	NR	None	NR	NR	NR
Type 2b							
Elfiky et al [43]	NR	NR	Probabilistic imputation	None	Grid search	4-fold CV	VI
Hanai et al [44]	Standardization	NR	NR	Between-variable correlation and PIM[j]	NR	5-fold CV	VI
Type 4							
Manz et al [37]	NR	NR	Constant value imputation	N/A	N/A	N/A	VI and coefficient
Karhade et al [36]	NR	NR	missForest multiple imputation	N/A	N/A	N/A	NR

[a]N/A: not applicable.

[b]NR: not reported.

[c]CV: cross-validation.

[d]VI: variable importance.

[e]LASSO: least absolute shrinkage and selection operator.

[f]LR: logistic regression.

[g]RF: random forest.

[h]PDP: partial dependence plot.

[i]LIME: local interpretable model-agnostic explanation.

[j]PIM: parameter-increasing method.

Solutions for Class Imbalance

All included studies reported that the mortality rate of their samples experienced some degree of class imbalance (Table 3). The median mortality rate was 20.0% (range 4%-56.2%), with 2.8 deaths in training samples per candidate predictor (range 0.5-12.3) in training samples. A type 1 study discussed the potential disadvantage of the issue and used a downsampling approach to handle imbalanced data. No information was provided on how the downsampling approach was conducted and its effectiveness on model performance in an unseen data set.

Sensitivity Analysis

Owing to the small number of included studies, we conducted a sensitivity analysis by including 1 study per research group to avoid the disproportionate effects of studies from a single group on our model performance and modeling practice evaluation. We observed similar issues concerning model development and evaluation practice after removing the studies by Manz et al [37] and Karhade et al [36,41]. For model performance, all algorithms still demonstrated good performance, with a median AUROC of 0.88 ranging from 0.81 to 0.89 (Multimedia Appendix 4 [36,37,41]). We detected changes in AUROC for all algorithms except RF and regularized LR (ranging from −0.008 to 0.065). Stochastic gradient boosting and support vector machine algorithms had the greatest changes in AUROC (ΔAUROC=0.06 and 0.065, respectively). However, the performance of these models in the sensitivity analysis may not be reliable as both algorithms were examined in the same study using a small sample (n=145).

Discussion

Principal Findings

Mortality prediction is a sensitive topic that, if done correctly, could assist with the provision of appropriate end-of-life care for patients with cancer. ML-based models have been developed to support the prediction; however, the current evidence has not yet been systematically examined. To fill this gap, we performed a systematic review evaluating 15 studies to summarize the evidence quality and the performance of ML-based models predicting short-term mortality for the identification of patients with cancer who may benefit from palliative care. Our findings suggest that the algorithms appeared to have promising overall discriminatory performance with respect to AUROC values, consistent with previous studies summarizing the performance of ML-based models supporting mortality predictions for other populations [16-19]. However, the results must be interpreted with caution because of the high ROB across the studies, as well as some evidence of the selective reporting of important performance metrics such as sensitivity and PPV, supporting previous studies reporting poor adherence to TRIPOD reporting items in ML studies [52]. We identified several common issues that could lead to biased models and misleading model performance estimates in the methods used to develop and evaluate the algorithms. The issues included the use of a single performance metric, incomplete reporting of or inappropriate data preprocessing and modeling, and small sample size. Further research is needed to establish a guideline for ML modeling, evaluation, and reporting to enhance the evidence quality in this area.

We found that the AUROC was predominantly used as the primary metric for model selection. Other performance metrics have been less discussed. However, the AUROC provides less information for determining whether the model is clinically beneficial, as it equally weighs sensitivity and specificity [53,54]. For instance, Manz et al [37] reported a model predicting 180-day mortality for patients with cancer with an AUROC of 0.89, showing the superior performance of the model [37]. However, their model demonstrated a low sensitivity of 0.27, indicating poor performance in identifying individuals at high risk of 180-day death. In practice, whether to stress sensitivity or specificity depends on the model's purpose. In the case of rare event prediction, we believe that sensitivity will usually be prioritized. Therefore, we strongly suggest that future studies report multiple discrimination metrics, including sensitivity, specificity, PPV, negative predictive value, F1 score, and the area under the precision–recall curve, to allow for a comprehensive evaluation [53-55].

We found no clear difference in performance between general and cancer-specific ML models for short-term mortality predictions (AUROC 0.87 for general models vs 0.86 for cancer-specific models). This finding aligns with a study reporting no performance benefit of disease-specific ML models over general ML models for hospital readmission predictions [56]. However, among the 15 included studies, 10 (67%) examined ML performance in short-term mortality for only a few types of cancer, which resulted in the ML in most cancer types remaining unexplored and compromising the comparison. In fact, a few disease-specific models examined in this review demonstrated exceptional performance and have the potential to provide disease-specific information to better guide clinical practice [40,47]. As such, we recommend that more research test ML models using various oncology-specific patient cohorts to predict short-term mortality to enable a full understanding of whether disease-specific ML models can bring advantages over limitations, such as higher development and implementation cost.

Only 33% (5/15) of the included studies compared their model with a traditional statistical model, such as univariate or multivariate LR [39,47,48,50,51]. Of the 15 studies, 1 (7%) reported that ML models were statistically more accurate, although all studies reported a superior AUROC of their ML models compared with statistical predictive models. This finding supports previous studies that reported that the performance benefit of ML over conventional modeling approaches is unclear at the current stage [57]. Thus, although we argue that the capacity of ML algorithms in dealing with nonlinear, high-dimensional data could benefit clinical practice by identifying additional risk factors for intervening to improve

patient outcomes beyond predictive performance, we encourage researchers to benchmark their ML models against conventional approaches to highlight the performance benefit of ML.

Our review suggests that the sample size consideration is missing for ML studies in the field, which is consistent with a previous review [58]. In fact, none of the included studies justified the appropriateness of their sample size, given the number of candidate predictors used in model development. Simulation studies have suggested that most ML modeling approaches require >200 data points related to the outcome per candidate predictor to reach a stable performance and mitigate optimistic models [59]. Unfortunately, none of the included studies met this criterion. Thus, we recommend that future studies justify the appropriateness of their sample size and use feature selection and dimensional reduction techniques before modeling to reduce the number of candidate predictors if a small sample is inevitably used.

Most studies used imbalanced data sets without additional procedures to address the issue, such as over- or downsampling. The effects of class-imbalanced data sets are unclear as sensitivity was often unreported and widely varied when it was reported. A study used a downsampling technique to balance their data set [38]. However, the authors did not report their model performance in a holdout validation data set. Thus, the effectiveness of this approach is unknown. Moreover, the effectiveness of other approaches, such as the synthetic minority oversampling technique [60], remains unexamined in this context. Further research is needed to examine whether these approaches can further improve the performance of ML models in predicting cancer mortality.

Most ML models predicting short-term cancer mortality were reported without intuitive interpretations of the prediction processes. It has been well-documented that ML acceptance by the larger medical community is limited because of the limited interpretability of ML-based models [53]. Despite the widespread use of variable importance analysis to reveal essential factors for the models in the included studies, it is unknown how the models used the factors to generate the predictions [61]. As the field progresses, global and local model interpretation approaches have been developed to explain ML models intuitively and visually at a data set and instance level [61]. The inclusion of these analyses to provide an intuitive model explanation may not only gain medical professionals' trust but also provide information guiding individualized care

plans and future investigations [62]. Therefore, we highly recommend that future studies *unbox* their models using various explanation analyses in addition to model performance.

Limitations

This review has several limitations. First, we did not quantitatively synthesize the model performance because of the clinical and methodological heterogeneity of the included studies. We believe that a meta-analysis of the model performance would provide clear evidence but should be conducted with enough homogeneous studies [63]. Second, the ROB of the studies may be inappropriately estimated because of the use of the prediction model ROB assessment tool checklist, which was developed for appraising predictive modeling studies using multivariable analysis. Some items may not apply, or additional items may be needed because of the differences in terminology, theoretical foundations, and procedures between ML-based and regression-based studies. Finally, the results of this review may be affected by reporting bias as we did not consider studies published outside of scientific journals or in non-English languages. Furthermore, our results could be compromised by the small number of included studies and the inclusion of studies by the same group (eg, 3 studies from Karhade et al [36,41,42]). However, we observed similar issues with model development and performance in our sensitivity analysis, suggesting that our evaluation likely reflects the current evidence in the literature. Despite these limitations, this review provides an overview of ML-based model performance in predicting short-term cancer mortality and leads to recommendations concerning model development and reporting.

Conclusions

In conclusion, we found signs of encouraging performance but also highlighted several issues concerning the way algorithms were trained, evaluated, and reported in the current literature. The overall ROB was high, and there was substantial uncertainty regarding the development and performance of the models in the real world because of incomplete reporting. Although some models are potentially clinically beneficial, we must conclude that none of the included studies produced an ML model that we considered suitable for clinical practice to support palliative care initiation and provision. We encourage further efforts to develop safe and effective ML models using modern standards of development and reporting.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Search strategy for all reference databases used.

XSL•FO

RenderX

[DOCX File , 24 KB - medinform_v10i3e33182_app1.docx]

Multimedia Appendix 2
Data collection tool.
[DOCX File , 26 KB - medinform_v10i3e33182_app2.docx]

Multimedia Appendix 3
Detailed record for the study selection process.
[XLSX File (Microsoft Excel File), 1276 KB - medinform_v10i3e33182_app3.xlsx]

Multimedia Appendix 4
Comparison of area under the receiver operating characteristic curve for each machine learning algorithm between the full and sensitivity analyses.
[DOCX File , 31 KB - medinform_v10i3e33182_app4.docx]

References

1. Earle CC, Neville BA, Landrum MB, Ayanian JZ, Block SD, Weeks JC. Trends in the aggressiveness of cancer care near the end of life. J Clin Oncol 2004;22(2):315-321. [doi: 10.1200/JCO.2004.08.136] [Medline: 14722041]
2. Tönnies J, Hartmann M, Jäger D, Bleyel C, Becker N, Friederich HC, et al. Aggressiveness of care at the end-of-life in cancer patients and its association with psychosocial functioning in bereaved caregivers. Front Oncol 2021;11:673147 [FREE Full text] [doi: 10.3389/fonc.2021.673147] [Medline: 34150639]
3. Ullgren H, Fransson P, Olofsson A, Segersvärd R, Sharp L. Health care utilization at end of life among patients with lung or pancreatic cancer. Comparison between two Swedish cohorts. PLoS One 2021;16(7):e0254673 [FREE Full text] [doi: 10.1371/journal.pone.0254673] [Medline: 34270589]
4. Bylicki O, Didier M, Riviere F, Margery J, Grassin F, Chouaid C. Lung cancer and end-of-life care: a systematic review and thematic synthesis of aggressive inpatient care. BMJ Support Palliat Care 2019;9(4):413-424 [FREE Full text] [doi: 10.1136/bmjspcare-2019-001770] [Medline: 31473652]
5. Jang TK, Kim DY, Lee SW, Park JY, Suh DS, Kim JH, et al. Trends in treatment during the last stages of life in end-stage gynecologic cancer patients who received active palliative chemotherapy: a comparative analysis of 10-year data in a single institution. BMC Palliat Care 2018;17(1):99 [FREE Full text] [doi: 10.1186/s12904-018-0348-7] [Medline: 30086748]
6. Wright AA, Keating NL, Ayanian JZ, Chrischilles EA, Kahn KL, Ritchie CS, et al. Family perspectives on aggressive cancer care near the end of life. JAMA 2016;315(3):284-292 [FREE Full text] [doi: 10.1001/jama.2015.18604] [Medline: 26784776]
7. Palliative and end-of-life care 2015-2016. National Quality Forum. 2016. URL: https://www.qualityforum.org/publications/2016/12/palliative_and_end-of-life_care_2015-2016.aspx [accessed 2022-02-28]
8. Avati A, Jung K, Harman S, Downing L, Ng A, Shah NH. Improving palliative care with deep learning. BMC Med Inform Decis Mak 2018;18(Suppl 4):122 [FREE Full text] [doi: 10.1186/s12911-018-0677-8] [Medline: 30537977]
9. Pirl WF, Lerner J, Traeger L, Greer JA, El-Jawahri A, Temel JS. Oncologists' dispositional affect and likelihood of end-of-life discussions. J Clin Oncol 2019;34(26_suppl):9 [FREE Full text] [doi: 10.1200/jco.2016.34.26_suppl.9]
10. Kale MS, Ornstein KA, Smith CB, Kelley AS. End-of-life discussions with older adults. J Am Geriatr Soc 2016;64(10):1962-1967 [FREE Full text] [doi: 10.1111/jgs.14285] [Medline: 27549494]
11. Obermeyer Z, Emanuel EJ. Predicting the future - big data, machine learning, and clinical medicine. N Engl J Med 2016;375(13):1216-1219 [FREE Full text] [doi: 10.1056/NEJMp1606181] [Medline: 27682033]
12. Sidey-Gibbons JA, Sidey-Gibbons CJ. Machine learning in medicine: a practical introduction. BMC Med Res Methodol 2019;19(1):64 [FREE Full text] [doi: 10.1186/s12874-019-0681-4] [Medline: 30890124]
13. Cruz JA, Wishart DS. Applications of machine learning in cancer prediction and prognosis. Cancer Inform 2007;2:59-77 [FREE Full text] [Medline: 19458758]
14. Kourou K, Exarchos TP, Exarchos KP, Karamouzis MV, Fotiadis DI. Machine learning applications in cancer prognosis and prediction. Comput Struct Biotechnol J 2014;13:8-17 [FREE Full text] [doi: 10.1016/j.csbj.2014.11.005] [Medline: 25750696]
15. Manz C, Parikh RB, Evans CN, Chivers C, Regli SB, Changolkar S, et al. Effect of integrating machine learning mortality estimates with behavioral nudges to increase serious illness conversions among patients with cancer: a stepped-wedge cluster randomized trial. J Clin Oncol 2020;38(15_suppl):12002 [FREE Full text] [doi: 10.1200/JCO.2020.38.15_suppl.12002]
16. Shillan D, Sterne JA, Champneys A, Gibbison B. Use of machine learning to analyse routinely collected intensive care unit data: a systematic review. Crit Care 2019;23(1):284 [FREE Full text] [doi: 10.1186/s13054-019-2564-9] [Medline: 31439010]
17. Rau CS, Kuo PJ, Chien PC, Huang CY, Hsieh HY, Hsieh CH. Mortality prediction in patients with isolated moderate and severe traumatic brain injury using machine learning models. PLoS One 2018;13(11):e0207192 [FREE Full text] [doi: 10.1371/journal.pone.0207192] [Medline: 30412613]

18. Bottino F, Tagliente E, Pasquini L, Di Napoli A, Lucignani M, Figà-Talamanca L, et al. COVID mortality prediction with machine learning methods: a systematic review and critical appraisal. J Pers Med 2021;11(9):893 [FREE Full text] [doi: 10.3390/jpm11090893] [Medline: 34575670]

19. Nardini C. Machine learning in oncology: a review. Ecancermedicalscience 2020;14:1065 [FREE Full text] [doi: 10.3332/ecancer.2020.1065] [Medline: 32728381]

20. Ramesh S, Chokkara S, Shen T, Major A, Volchenboum SL, Mayampurath A, et al. Applications of artificial intelligence in pediatric oncology: a systematic review. JCO Clin Cancer Inform 2021;5:1208-1219 [FREE Full text] [doi: 10.1200/CCI.21.00102] [Medline: 34910588]

21. Wang W, Kiik M, Peek N, Curcin V, Marshall IJ, Rudd AG, et al. A systematic review of machine learning models for predicting outcomes of stroke with structured data. PLoS One 2020;15(6):e0234722 [FREE Full text] [doi: 10.1371/journal.pone.0234722] [Medline: 32530947]

22. Lazer D, Kennedy R, King G, Vespignani A. Big data. The parable of Google flu: traps in big data analysis. Science 2014;343(6176):1203-1205. [doi: 10.1126/science.1248506] [Medline: 24626916]

23. Obermeyer Z, Powers B, Vogeli C, Mullainathan S. Dissecting racial bias in an algorithm used to manage the health of populations. Science 2019;366(6464):447-453. [doi: 10.1126/science.aax2342] [Medline: 31649194]

24. Pfob A, Mehrara BJ, Nelson JA, Wilkins EG, Pusic AL, Sidey-Gibbons C. Towards patient-centered decision-making in breast cancer surgery: machine learning to predict individual patient-reported outcomes at 1-year follow-up. Ann Surg (forthcoming) 2021. [doi: 10.1097/SLA.0000000000004862] [Medline: 33914464]

25. Roberts M, Driggs D, Thorpe M, Gilbey J, Yeung M, Ursprung S, AIX-COVNET. Common pitfalls and recommendations for using machine learning to detect and prognosticate for COVID-19 using chest radiographs and CT scans. Nat Mach Intell 2021;3(3):199-217 [FREE Full text] [doi: 10.1038/s42256-021-00307-0]

26. Stupple A, Singerman D, Celi LA. The reproducibility crisis in the age of digital medicine. NPJ Digit Med 2019;2:2 [FREE Full text] [doi: 10.1038/s41746-019-0079-z] [Medline: 31304352]

27. He J, Baxter SL, Xu J, Xu J, Zhou X, Zhang K. The practical implementation of artificial intelligence technologies in medicine. Nat Med 2019;25(1):30-36 [FREE Full text] [doi: 10.1038/s41591-018-0307-0] [Medline: 30617336]

28. McDermott MB, Wang S, Marinsek N, Ranganath R, Foschini L, Ghassemi M. Reproducibility in machine learning for health research: still a ways to go. Sci Transl Med 2021;13(586):eabb1655. [doi: 10.1126/scitranslmed.abb1655] [Medline: 33762434]

29. Uddin MF. Addressing accuracy paradox using enhanced weighted performance metric in machine learning. In: Sixth HCT Information Technology Trends. 2019 Presented at: ITT '19; November 20-21, 2019; Ras Al Khaimah, UAE p. 319-324 URL: https://doi.org/10.1109/itt48889.2019.9075071 [doi: 10.1109/itt48889.2019.9075071]

30. Valverde-Albacete FJ, Peláez-Moreno C. 100% classification accuracy considered harmful: the normalized information transfer factor explains the accuracy paradox. PLoS One 2014;9(1):e84217 [FREE Full text] [doi: 10.1371/journal.pone.0084217] [Medline: 24427282]

31. Aromataris E, Munn Z. JBI Manual for Evidence Synthesis. Adelaide, Australia: Joanna Briggs Institute; 2020.

32. Page MJ, McKenzie JE, Bossuyt PM, Boutron I, Hoffmann TC, Mulrow CD, et al. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. BMJ 2021;372:n71 [FREE Full text] [doi: 10.1136/bmj.n71] [Medline: 33782057]

33. Moons KG, Wolff RF, Riley RD, Whiting PF, Westwood M, Collins GS, et al. PROBAST: a tool to assess risk of bias and applicability of prediction model studies: explanation and elaboration. Ann Intern Med 2019;170(1):W1-33 [FREE Full text] [doi: 10.7326/M18-1377] [Medline: 30596876]

34. Collins GS, Reitsma JB, Altman DG, Moons KG. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): the TRIPOD statement. Ann Intern Med 2015;162(1):55-63 [FREE Full text] [doi: 10.7326/M14-0697] [Medline: 25560714]

35. Oh SE, Seo SW, Choi MG, Sohn TS, Bae JM, Kim S. Prediction of overall survival and novel classification of patients with gastric cancer using the survival recurrent network. Ann Surg Oncol 2018;25(5):1153-1159. [doi: 10.1245/s10434-018-6343-7] [Medline: 29497908]

36. Karhade AV, Ahmed A, Pennington Z, Chara A, Schilling A, Thio QC, et al. External validation of the SORG 90-day and 1-year machine learning algorithms for survival in spinal metastatic disease. Spine J 2020;20(1):14-21. [doi: 10.1016/j.spinee.2019.09.003] [Medline: 31505303]

37. Manz CR, Chen J, Liu M, Chivers C, Regli SH, Braun J, et al. Validation of a machine learning algorithm to predict 180-day mortality for outpatients with cancer. JAMA Oncol 2020;6(11):1723-1730 [FREE Full text] [doi: 10.1001/jamaoncol.2020.4331] [Medline: 32970131]

38. Sena GR, Lima TP, Mello MJ, Thuler LC, Lima JT. Developing machine learning algorithms for the prediction of early death in elderly cancer patients: usability study. JMIR Cancer 2019;5(2):e12163 [FREE Full text] [doi: 10.2196/12163] [Medline: 31573896]

39. Parikh RB, Manz C, Chivers C, Regli SH, Braun J, Draugelis ME, et al. Machine learning approaches to predict 6-month mortality among patients with cancer. JAMA Netw Open 2019;2(10):e1915997 [FREE Full text] [doi: 10.1001/jamanetworkopen.2019.15997] [Medline: 31651973]

40. Zhang M, Yin F, Chen B, Li B, Li YP, Yan LN, et al. Mortality risk after liver transplantation in hepatocellular carcinoma recipients: a nonlinear predictive model. Surgery 2012;151(6):889-897. [doi: 10.1016/j.surg.2011.12.034] [Medline: 22341043]

41. Karhade AV, Thio QC, Ogink PT, Shah AA, Bono CM, Oh KS, et al. Development of machine learning algorithms for prediction of 30-day mortality after surgery for spinal metastasis. Neurosurgery 2019;85(1):E83-E91. [doi: 10.1093/neuros/nyy469] [Medline: 30476188]

42. Karhade AV, Thio QC, Ogink PT, Bono CM, Ferrone ML, Oh KS, et al. Predicting 90-day and 1-year mortality in spinal metastatic disease: development and internal validation. Neurosurgery 2019;85(4):E671-E681. [doi: 10.1093/neuros/nyz070] [Medline: 30869143]

43. Elfiky AA, Pany MJ, Parikh RB, Obermeyer Z. Development and application of a machine learning approach to assess short-term mortality risk among patients with cancer starting chemotherapy. JAMA Netw Open 2018;1(3):e180926 [FREE Full text] [doi: 10.1001/jamanetworkopen.2018.0926] [Medline: 30646043]

44. Hanai T, Yatabe Y, Nakayama Y, Takahashi T, Honda H, Mitsudomi T, et al. Prognostic models in patients with non-small-cell lung cancer using artificial neural networks in comparison with logistic regression. Cancer Sci 2003;94(5):473-477 [FREE Full text] [doi: 10.1111/j.1349-7006.2003.tb01467.x] [Medline: 12824896]

45. Nilsaz-Dezfouli H, Abu-Bakar MR, Arasan J, Adam MB, Pourhoseingholi MA. Improving gastric cancer outcome prediction using single time-point artificial neural network models. Cancer Inform 2017;16:1176935116686062 [FREE Full text] [doi: 10.1177/1176935116686062] [Medline: 28469384]

46. Arostegui I, Gonzalez N, Fernández-de-Larrea N, Lázaro-Aramburu S, Baré M, Redondo M, REDISSEC CARESS-CCR Group. Combining statistical techniques to predict postsurgical risk of 1-year mortality for patients with colon cancer. Clin Epidemiol 2018;10:235-251 [FREE Full text] [doi: 10.2147/CLEP.S146729] [Medline: 29563837]

47. Biglarian A, Hajizadeh E, Kazemnejad A, Zali M. Application of artificial neural network in predicting the survival rate of gastric cancer patients. Iran J Public Health 2011;40(2):80-86 [FREE Full text] [Medline: 23113076]

48. Klén R, Salminen AP, Mahmoudian M, Syvänen KT, Elo LL, Boström PJ. Prediction of complication related death after radical cystectomy for bladder cancer with machine learning methodology. Scand J Urol 2019;53(5):325-331. [doi: 10.1080/21681805.2019.1665579] [Medline: 31552774]

49. Chiu HC, Ho TW, Lee KT, Chen HY, Ho WH. Mortality predicted accuracy for hepatocellular carcinoma patients with hepatic resection using artificial neural network. ScientificWorldJournal 2013;2013:201976 [FREE Full text] [doi: 10.1155/2013/201976] [Medline: 23737707]

50. Bertsimas D, Kung J, Trichakis N, Wang Y, Hirose R, Vagefi PA. Development and validation of an optimized prediction of mortality for candidates awaiting liver transplantation. Am J Transplant 2019;19(4):1109-1118 [FREE Full text] [doi: 10.1111/ajt.15172] [Medline: 30411495]

51. Shi HY, Lee KT, Wang JJ, Sun DP, Lee HH, Chiu CC. Artificial neural network model for predicting 5-year mortality after surgery for hepatocellular carcinoma: a nationwide study. J Gastrointest Surg 2012;16(11):2126-2131. [doi: 10.1007/s11605-012-1986-3] [Medline: 22878787]

52. Dhiman P, Ma J, Navarro CA, Speich B, Bullock G, Damen JA, et al. Reporting of prognostic clinical prediction models based on machine learning methods in oncology needs to be improved. J Clin Epidemiol 2021;138:60-72 [FREE Full text] [doi: 10.1016/j.jclinepi.2021.06.024] [Medline: 34214626]

53. El Naqa I, Ruan D, Valdes G, Dekker A, McNutt T, Ge Y, et al. Machine learning and modeling: data, validation, communication challenges. Med Phys 2018;45(10):e834-e840 [FREE Full text] [doi: 10.1002/mp.12811] [Medline: 30144098]

54. Johnston SS, Fortin S, Kalsekar I, Reps J, Coplan P. Improving visual communication of discriminative accuracy for predictive models: the probability threshold plot. JAMIA Open 2021;4(1):ooab017 [FREE Full text] [doi: 10.1093/jamiaopen/ooab017] [Medline: 33733059]

55. Halligan S, Altman DG, Mallett S. Disadvantages of using the area under the receiver operating characteristic curve to assess imaging tests: a discussion and proposal for an alternative approach. Eur Radiol 2015;25(4):932-939 [FREE Full text] [doi: 10.1007/s00330-014-3487-0] [Medline: 25599932]

56. Sutter T, Roth JA, Chin-Cheong K, Hug BL, Vogt JE. A comparison of general and disease-specific machine learning models for the prediction of unplanned hospital readmissions. J Am Med Inform Assoc 2021;28(4):868-873 [FREE Full text] [doi: 10.1093/jamia/ocaa299] [Medline: 33338231]

57. Christodoulou E, Ma J, Collins GS, Steyerberg EW, Verbakel JY, Van Calster B. A systematic review shows no performance benefit of machine learning over logistic regression for clinical prediction models. J Clin Epidemiol 2019;110:12-22. [doi: 10.1016/j.jclinepi.2019.02.004] [Medline: 30763612]

58. Ben-Israel D, Jacobs WB, Casha S, Lang S, Ryu WH, de Lotbiniere-Bassett M, et al. The impact of machine learning on patient care: a systematic review. Artif Intell Med 2020;103:101785. [doi: 10.1016/j.artmed.2019.101785] [Medline: 32143792]

59. van der Ploeg T, Austin PC, Steyerberg EW. Modern modelling techniques are data hungry: a simulation study for predicting dichotomous endpoints. BMC Med Res Methodol 2014;14:137 [FREE Full text] [doi: 10.1186/1471-2288-14-137] [Medline: 25532820]

60.   Chawla NV, Bowyer KW, Hall LO, Kegelmeyer WP. SMOTE: synthetic minority over-sampling technique. J Artif Intell Res 2002;16(1):321-357. [doi: 10.1613/jair.953]
61.   Molnar C. Interpretable Machine Learning: A Guide for Making Black Box Models Explainable. Morrisville, NC: Lulu Publishing; 2019.
62.   Li R, Shinde A, Liu A, Glaser S, Lyou Y, Yuh B, et al. Machine learning-based interpretation and visualization of nonlinear interactions in prostate cancer survival. JCO Clin Cancer Inform 2020;4:637-646 [FREE Full text] [doi: 10.1200/CCI.20.00002] [Medline: 32673068]
63.   Higgins JP, Thomas J, Chandler J, Cumpston M, Li T, Page MJ, et al. Cochrane Handbook for Systematic Reviews of Interventions. 2nd Edition. Chichester, UK: John Wiley & Sons; 2019.

Abbreviations

AUROC: area under the receiver operating characteristic curve
EHR: electronic health record
LR: logistic regression
ML: machine learning
PPV: positive predictive value
PRISMA: Preferred Reporting Items for Systematic Reviews and Meta-Analyses
PROSPERO: International Prospective Register of Systematic Reviews
RF: random forest
ROB: risk of bias
TRIPOD: the transparent reporting of a multivariable prediction model for individual prognosis or diagnosis

Review

The Role of Health Kiosks: Scoping Review

Inocencio Daniel Maramba[1], BSc, MSc, MD; Ray Jones[1], BSc, MSc, PhD; Daniela Austin[1], BSc, MPsych; Katie Edwards[1], BSc, MSc; Edward Meinert[1], MA, MSc, MBA, MPA, PhD; Arunangsu Chatterjee[1], BEng, MSc, PCHE, PhD

Centre for Health Technology, University of Plymouth, Plymouth, United Kingdom

Corresponding Author:
Inocencio Daniel Maramba, BSc, MSc, MD
Centre for Health Technology
University of Plymouth
Drake Circus
Plymouth, PL4 8AA
United Kingdom
Phone: 44 1752 587484
Email: inocencio.maramba@plymouth.ac.uk

Abstract

Background: Health kiosks are publicly accessible computing devices that provide access to services, including health information provision, clinical measurement collection, patient self–check-in, telemonitoring, and teleconsultation. Although the increase in internet access and ownership of smart personal devices could make kiosks redundant, recent reports have predicted that the market will continue to grow.

Objective: We seek to clarify the current and future roles of health kiosks by investigating the settings, roles, and clinical domains in which kiosks are used; whether usability evaluations of health kiosks are being reported, and if so, what methods are being used; and what the barriers and facilitators are for the deployment of kiosks.

Methods: We conducted a scoping review using a bibliographic search of Google Scholar, PubMed, and Web of Science databases for studies and other publications between January 2009 and June 2020. Eligible papers described the implementation as primary studies, systematic reviews, or news and feature articles. Additional reports were obtained by manual searching and querying the key informants. For each article, we abstracted settings, purposes, health domains, whether the kiosk was opportunistic or integrated with a clinical pathway, and whether the kiosk included usability testing. We then summarized the data in frequency tables.

Results: A total of 141 articles were included, of which 134 (95%) were primary studies, and 7 (5%) were reviews. Approximately 47% (63/134) of the primary studies described kiosks in secondary care settings. Other settings included community (32/134, 23.9%), primary care (24/134, 17.9%), and pharmacies (8/134, 6%). The most common roles of the health kiosks were providing health information (47/134, 35.1%), taking clinical measurements (28/134, 20.9%), screening (17/134, 12.7%), telehealth (11/134, 8.2%), and patient registration (8/134, 6.0%). The 5 most frequent health domains were multiple conditions (33/134, 24.6%), HIV (10/134, 7.5%), hypertension (10/134, 7.5%), pediatric injuries (7/134, 5.2%), health and well-being (6/134, 4.5%), and drug monitoring (6/134, 4.5%). Kiosks were integrated into the clinical pathway in 70.1% (94/134) of studies, opportunistic kiosks accounted for 23.9% (32/134) of studies, and in 6% (8/134) of studies, kiosks were used in both. Usability evaluations of kiosks were reported in 20.1% (27/134) of papers. Barriers (e.g., use of expensive proprietary software) and enablers (e.g., handling of on-demand consultations) of deploying health kiosks were identified.

Conclusions: Health kiosks still play a vital role in the health care system, including collecting clinical measurements and providing access to web-based health services and information to those with little or no digital literacy skills and others without personal internet access. We identified research gaps, such as training needs for teleconsultations and scant reporting on usability evaluation methods.

KEYWORDS

Introduction

Rationale

Health kiosks are publicly accessible computing devices used to provide access to a variety of services in the health care system. In a 2009 review, Jones [1] classified health kiosks as (1) opportunistic, placed in locations and waiting for use, and (2) integrated, designed into the clinical process. Seven possible roles for health kiosks were identified: taking medical histories, health promotion, self-assessment, consumer feedback, patient registration, patient access to records, and remote consultations.

At that time, 65% of households in the United Kingdom had internet access. By 2020, internet access had increased to 96% of households, most (98%) with a fixed broadband connection and 64% of households having internet access through mobile devices [2]. Older people have started to close the digital gap with younger age groups: recent internet use (the preceding 3 months) increased from 52% to 83% among individuals aged 65 to 74 years and from 20% to 47% among adults aged ≥75 years from between 2011 and 2019 [3]. Smartphone and tablet ownership in the United Kingdom has increased from 26% and 2% in 2011 to 78% and 58% in 2018, respectively [4]. These trends were also reflected worldwide. For example, 318,000 health-related apps for smartphones and tablets were listed in the app stores as of 2019 [5].

Data from the International Telecommunication Union show that these trends are reflected worldwide:

- The percentage of the world population with access to the internet increased from 26% (1.8 billion people) in 2009 to 51% (4 billion people) in 2019, broken down regionally as follows: 7% to 6% to 28.6% in Africa, 20.6% to 54.6% in the Arab States, 19% to 44.5% in Asia and the Pacific, 24.3% to 72.8% in the Commonwealth of Independent States, 59.6% to 82.5% in Europe, and 46.3% to 76.7% in the Americas.
- The number of mobile phone subscriptions per 100 people worldwide increased from 68 in 2009 to 107.8 in 2019 (meaning that in 2019 some people had more than one subscription).
- Fixed broadband connections per 100 people worldwide increased from 6.9 in 2009 to 14.8 in 2019. [6].

However, these developments do not make health kiosks redundant.

Despite these trends, various authors predict continued and even growing use of kiosks. Chen [7] has predicted that telehealth kiosks will be widespread by 2023. Similarly, a recent blog piece by Kochelek [8] has stated that health kiosks will be essential in the changing medical landscape for the following reasons: (1) kiosks will streamline patient check-in; (2) human-to-human contact will be minimized by the use of kiosks, which is vital during the coronavirus pandemic; and (3) telehealth kiosks placed in private areas of strategic locations will provide access to patient care for the public, and kiosks in group homes can also provide care for individuals who are immune compromised, reducing the need for travel and the risk of exposure. Thus, in contrast to expectations of the death of kiosks because of the use of mobile technologies, an alternative view is that health kiosks will still be a major part of the digital health landscape in the foreseeable future.

With the above in mind, we saw the need to investigate the evolution of the roles of health kiosks in the past decade and what possible roles they may play in the future. We were aware that there may have been reviews of health kiosks published since the work of Jones [1] in 2009, and an investigation by one of the authors revealed that the latest review before starting this one was published in 2013. As there have been great changes technologically in the past 7 years, the authors believed conducting a new review of the literature about health kiosks was justified.

Background

Health Kiosks Versus Personal Smart Devices

As mentioned previously, the roles played by kiosks a decade ago may now be performed by personal smart devices (smartphones and tablets), especially in the delivery of health information. In 2019, 79% of adults (aged ≥18 years) in the United Kingdom owned a smartphone, and tablet ownership was estimated at 58%. However, this is subject to age differences, as only 40% of adults aged ≥65 years own smartphones [9,10]. Thus, health kiosks still play a role in providing access to health services to this segment of the population.

Even for smartphone and tablet owners, health information delivery via kiosks may still be useful as the information can be tailored, vetted, and delivered at the point of service. Although this may also be possible through smartphone apps, the app would need to be properly accredited and evaluated for accuracy, and the user would need to download it to their phone for it to be useful. However, tailored and vetted information delivered by a kiosk is already available without any further action on the part of the user.

The collection of clinical measurements is where health kiosks currently outperform personal smart devices. Although there are clinical measurement devices that can be connected to smartphones and tablets, such as blood pressure (BP) monitors, heart rate trackers, and glucose monitors, they have not yet become widespread in use. Health kiosks with linked measurement devices, such as stethoscopes, otoscopes, dermatoscopes, pulse oximeters, and BP monitors, can collect clinical data for telemonitoring or synchronous teleconsultations.

Health Kiosks for Remote Consultation

Overview

Teleconsultations are now also possible on smart devices or PCs without the need for a health kiosk. As reported in the news, during the lockdown period caused by the COVID-19 pandemic of 2020, only 7 of 100 general practitioner (GP) consultations were performed face to face, with the rest being done remotely. However, it is interesting to note that most of these consultations were still being conducted through telephone or text [11,12]. The news article also stated that there were still situations where patients needed to attend a practice in person, such as when BP or oxygen saturation needed to be read. These readings can be

obtained using a properly equipped health kiosk. In a way, this is analogous to the existing situation of web-based banking apps and cash machines. The availability of web-based banking has not done away with the need for cash points, and banks have not yet relegated them to the scrap heap. Although web-based consultations can be facilitated through mobile devices, a substantial number of such encounters will require some physical examination or measurement using diagnostic instruments, which health kiosks can provide in lieu of face-to-face consultations. Manufacturers now provide solutions where health kiosks could be reconceptualized as *health pods*, similar to *photo booths*, with a private space to have consultations along with a range of devices performing point-of-care clinical measurements (eg, BP monitors, pulse oximeters, and stethoscopes). This is particularly useful in rural, remote, and deprived communities. There are several telehealth kiosk products currently on offer that follow this model. Some examples of these are the kiosks offered by MedicSpot, Amwell, RPM Solutions, and H4D.

MedicSpot is a web-based GP service in the United Kingdom that allows patients to connect to a physician via kiosks placed in pharmacies. It is available at ≥300 locations across the United Kingdom. The kiosk is available for walk-in consultations without appointments and contains medical equipment for examinations. The service provides patients access to a connected stethoscope; pulse oximeter; BP monitor; contactless thermometer; and an inspection camera to check the ear, nose, and throat. This is a private service that charges £39 (US $51.70) per consultation. MedicSpot has recently partnered with the British supermarket chain Asda to offer in-store GP video consultations with diagnostics [13-16].

The kiosk line of Amwell, which is based in Massachusetts, United States, comprises a fully enclosed kiosk model, freestanding open console kiosk, and tabletop kiosk model. All models include a touchscreen interface, integrated camera, credit card reader, handset for private audio, and sanitation features. They can be equipped with biometric and clinical measurement devices that allow virtual monitoring of a patient's vital signs in real time. These include stethoscopes, otoscopes, pulse oximeters, BP cuffs, dermatoscopes, and thermometers [17]. Signs of Amwell's growing strength in the telehealth market include a report that the company would be going public later in 2020, as well as raising US $194 million in funding by May 2020 [18].

Meanwhile, H4D, a health technology start-up based in Paris, France, completed a €15 million (US$ 16.4 million) round of funding in June 2020. H4D developed a telemedicine platform centered on the Consult Station, which is a connected telemedicine booth. It comprises all the necessary instruments and sensors for physicians to consult with patients via videoconference. The Consult Station has been deployed to ensure continuity of care and treatment for patients who are chronically ill and cannot be safely treated in traditional health care facilities.

It is worth noting that the abovementioned implementations were all in the private health sectors of the United Kingdom, the United States, and France. The adoption of health kiosks for teleconsultation by government-run health systems has been slow because of the strict rules for suppliers of equipment. Publicly funded health systems require evidence from numerous trials before adopting new technologies.

Health Kiosks for Responding to the COVID-19 Pandemic

Health authorities such as the World Health Organization and the Centers for Disease Control and Prevention have strongly urged ways of minimizing physical contact between patients and health care providers, otherwise known as *medical distancing*. Telehealth services are rapidly becoming one of the primary methods of reducing health care–related COVID-19 transmissions and protecting health personnel [19]. Telehealth kiosks equipped with monitoring and clinical measurement devices will allow comprehensive medical examination of the patient while maintaining medical distancing. The need for medical distancing is one of the drivers of the increased adoption of telemedicine kiosks.

In response to the COVID-19 pandemic, Elephant Kiosks (Cornwall, United Kingdom) introduced the COVID-19 Reception Kiosk, which offers the first point of contact for visitors and staff in workplaces, care homes, schools, and other public places. It offers an integrated contactless temperature check, a COVID-19 questionnaire, and email alerts to managers or the reception. It meets the infection control guidance and can be used to support contact tracing [20].

The H4d Consult Station has also been used to support hospitals during the COVID-19 pandemic, notably the Ramsay Health Vert-Galant Hospital's emergency department (ED). The station was used to provide an initial screen and detect suspected COVID-19 cases. Using the Consult Station, the hospital was able to substantially reduce nurses' intake time and protect them from the virus [21,22].

Health Kiosks for Remote and Rural Locations

One of the benefits of telehealth kiosks is making medical and specialist care available to remote places that medical professionals rarely visit. These places can be remote rural areas with poor infrastructure in countries such as India and Canada [23-25] or geographically remote places such as island communities or offshore installations, such as Scotland [26].

In their study, Nachum et al [27] found that in the United States, those who used teleconsultation kiosks were significantly more likely to be visitors to the area rather than local people, suggesting that a visit to the kiosk represented an opportunity to access care when not familiar with local services. This could suggest that the implementation of kiosks in areas experiencing high levels of tourism could help with their impact on health care provision. For example, the remote region of Cornwall, located on the southwesterly peninsula of the United Kingdom, sees as many as 4 million tourism trips each year, predominantly in the summer, putting huge pressure on infrastructure, including health care services [28]. In 2004, the estimated cost of the provision of primary health care to nonresidents in Cornwall was £4.7 million (US $6.23 million) [29]. The deployment of teleconsultation kiosks to cater to nonresidents could ease the pressure on local health services, especially if less urgent

conditions could be managed by an autonomous mode of operation, with more urgent cases being seen *live* by a remote health care professional.

Objective

To clarify how the role of health kiosks has evolved in the past decade and what roles they may play in the future, we conducted a scoping review. The primary objectives of this review are to describe the scope of kiosk use in health care (by patients, health care providers, or the general public), examine the roles played by health kiosks in the health care system, and investigate the barriers to and facilitators of the deployment of kiosks. We have developed the following research questions to address these objectives:

- What are the settings and health domains in which health kiosks are deployed, and what health services are they delivering?
- Are health kiosk interventions evaluated for usability, which has been identified as being important for effective digital health [30-32]?
- Finally, what are the barriers to and facilitators of the deployment of kiosks, especially for teleconsultation (eg, resources, infrastructure, and training [33])?

Methods

Overview

A scoping review is defined as a type of research synthesis that aims to "map the literature on a particular topic or research area" [34,35]. We undertook a scoping review of the published literature, as well as the gray literature available from websites and social media. To ensure that this was comprehensive, we also identified key informants from the contacts database of the Ehealth Productivity and Innovation in Cornwall and the Isles of Scilly Project [36], and through a Google search, we gathered information from them via emails and video calls.

Definition of Health Kiosk

Computerized health kiosks have been defined as "freestanding units containing computer programs that provide users with information or services." [37]. For this review, we used the following definition of health kiosks: public access computing devices providing or collecting information at any point in the health care journey. Kiosks are normally owned by a health service provider but used by various members of the public. Health application software (*apps*) were only included if they were made available on a public access device; if they were installed on personally owned devices such as smartphones, tablets, laptops, and desktop computers, they were excluded.

Study Eligibility

Articles were included if they met the following criteria:

- Were about an actual implementation of a health kiosk and not a specification or nonfunctional prototype
- Were published in peer-reviewed publications, trade publications, and web-based health information technology publications
- Were published in the English language

- Were published between January 1, 2009, and June 1, 2020; we chose this period to update the previous review by Jones [1], which was published in 2009

Articles were excluded if they were design proposals for kiosks or nonfunctioning prototypes, if the device was a personal smart device rather than a publicly accessible device, or if they were in a language other than English.

Information Sources and Search Terms

The first source was published in the literature. We searched three electronic literature databases: Web of Science, PubMed (including MEDLINE), and Google Scholar.

The primary search term was *health kiosk*, which we used for all 3 databases. We trialed using the search terms *[health]AND [kiosk] AND[touchscreen]*

As used in previous reviews, this resulted in the inclusion of papers mostly about personal smart devices such as smartphones and tablets, which we did not classify as kiosks.

The final search terms were as follows:

- PubMed:
  ((health[MeSH Terms] OR health[All Fields] OR health s[All Fields] OR "healthful"[All Fields] OR healthfulness[All Fields] OR healths[All Fields]) AND (kiosk[All Fields] OR kiosks[All Fields])) AND ((2009/1/1:2020/6/1[pdat]) AND (english[Filter]))
- Web of Science (advanced search):
  ALL=health AND ALL=kiosk
- Google Scholar (advanced search): exact phrase health kiosk
  anywhere in the article between 2009 and 2020

Gray literature and social media were also searched using the Google search engine for reports and publications on relevant websites, as well as the search function on two social media websites: Facebook and Twitter. Key informants (kiosk manufacturers) were identified through a Google search and the contact database of the Ehealth Productivity and Innovation in Cornwall and the Isles of Scilly Project; they were contacted via email and video calls. We asked the manufacturers about the use cases of their kiosk offerings, training needs for kiosk use, barriers and facilitators for successful deployment, and any relevant publications. A total of 3 kiosk manufacturers from around the world responded to our inquiries.

Study Selection

We collated citations from the literature search using the Mendeley (Elsevier) reference management software, and duplicate citations were eliminated. Author IDM screened the titles and abstracts to determine whether the study met the inclusion criteria. The studies were classified as either included or excluded. All articles classified as *included* had their full text retrieved for further review. DA, KE, and IDM then independently evaluated the full text of each study according to the agreed inclusion criteria. Disagreements were resolved by voting, with the third member serving as the tiebreaker. Critical appraisal was not performed as we did not compare study results, and streamlining of methods is acceptable in a

scoping review [35]. We have presented the search results in a PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) flow diagram (Figure 1). A total of 141 articles met the inclusion criteria after a full-text review.

Figure 1. Diagram of articles reviewed for inclusion.
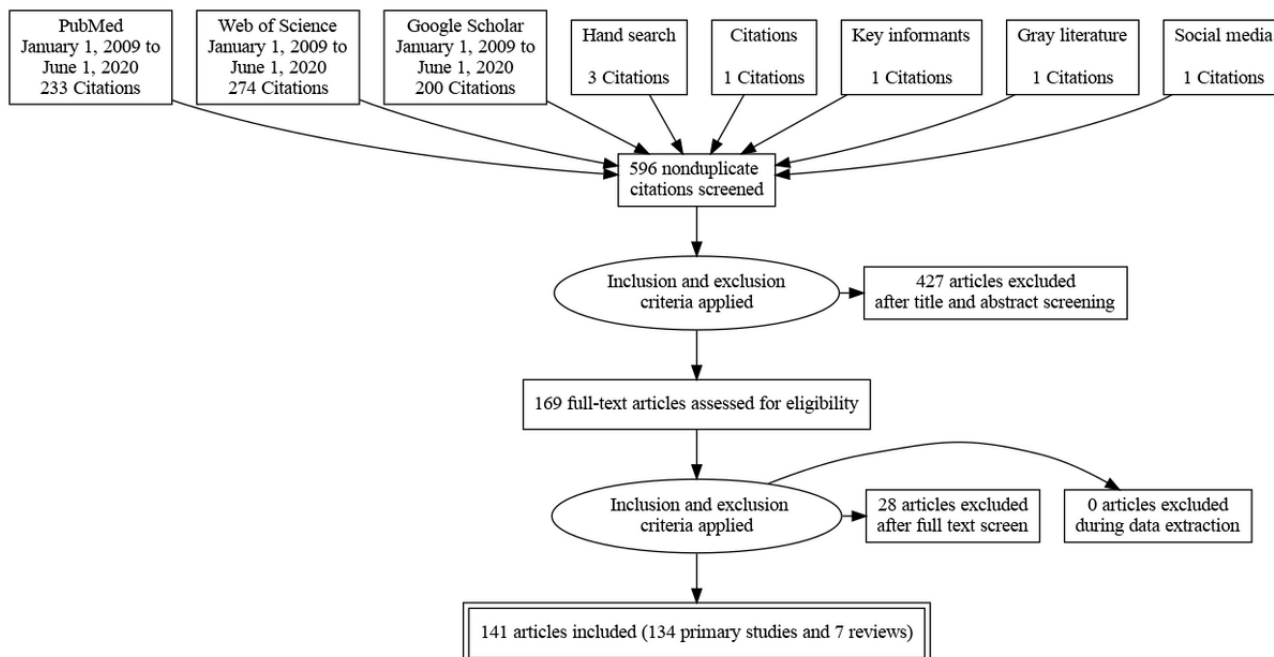
Data Extraction and Analysis

A data extraction form was created based on the table of published studies on health kiosks used in the paper by Jones [1]. The extracted data items included the setting, number of kiosks, year of publication, country of implementation, type of access to the kiosk (opportunistic or referred), purpose of the kiosk, health conditions targeted by the kiosk, and whether and how the kiosk was evaluated for usability. Other significant information about the kiosk study was included as comments. DA, KE, and IDM performed the data extraction. The results were then encoded into a Microsoft Excel spreadsheet. IDM rechecked the data extraction table for consistency, with differences in coding resolved through discussions among IDM, DA, and KE. Frequency tables and graphs were constructed using R (version 4.2.0) [38].

Results

Overview

We present the results of our literature search as follows: (1) settings, purposes, and conditions addressed by the kiosks in the included papers; (2) country of publication; (3) year of publication; (4) type of kiosk access; (5) patient self–check-in kiosks; (6) reporting on the usability evaluation of kiosks; (7) telemonitoring and teleconsultation kiosks, training needs, and barriers to and enablers of adoption.

Identified Publications

We identified 141 publications (Multimedia Appendix 1 [1,39-126]) by searching the PubMed (MEDLINE), Web of Science, and Google Scholar databases (Figure 1). All but 5% (7/141) were primary articles describing health kiosk implementations in clinical or community settings. Of the 7 systematic reviews, 3 (43%) were general reviews [1,39,40], 3 (43%) reviewed kiosks used for particular purposes (health information) [41-43], and 1 (14%) reviewed studies on kiosks used for BP monitoring [44]. The characteristics of the 141 included studies are summarized in Multimedia Appendix 1.

Settings, Purposes, and Conditions

In the 134 primary studies, the most frequent setting (n=61, 47%) was secondary care, which was subdivided into specialty and outpatient clinics (n=34, 54%), EDs (n=26, 43%), and inpatient settings (n=5, 8%). The most frequently cited purpose (45/134, 33.6%) was providing health information (Table 1). Kiosk implementation most frequently targeted multiple health domains or conditions, followed by HIV. The setting *specialty clinics* included clinics such as sexual health and cancer clinics, where patients are referred from primary care and hospital department outpatient clinics. EDs are acute care centers, including accident and EDs within hospitals. Primary care settings included general practices, family medicine clinics, and community clinics. Community refers to the settings in which kiosks were deployed in nonclinical venues, including churches [45,46] and community centers [45,47]. *Multiple* refers to the implementation of kiosks in multiple categories; for example, in both a community pharmacy (retail outlet for medications and other health care–related products) and a library [48] or simultaneously in a nonclinical (eg, a social service agency, a church, a school, and a coffee shop) and a clinical (primary care clinic) setting [49].

Table 1. Summary of settings, purposes, and health domains for the primary studies (N=134).

Categories	Values, n (%)
Settings	
Secondary care	63 (47)
Specialty clinic	34 (54)
Emergency department	26 (41)
Hospital inpatient	5 (8)
Community	32 (23.9)
Primary care	24 (17.9)
Pharmacy	8 (6)
Multiple	7 (5.2)
Purposes	
Health information	47 (35.1)
Clinical measurements	28 (20.9)
Screening	17 (12.7)
Telehealth	11 (8.2)
Patient registration	8 (6)
Patient feedback	6 (4.5)
Medication adherence	6 (4.5)
Patient outcomes data	5 (3.7)
Other	3 (2.2)
Patient triage	3 (2.2)
Health domains	
Multiple conditions	33 (24.6)
HIV	10 (7.5)
Hypertension	10 (7.5)
Pediatric injuries	7 (5.2)
Health and well-being	6 (4.5)
Medication	6 (4.5)
Cardiovascular disease	5 (3.7)
Mental health	4 (3)
Sexual health	4 (3)
Acute care—emergency department	3 (2.2)
Dementia	3 (2.2)
Others	43 (32.1)

Table 2 shows the purposes of the kiosks arranged according to the setting. The most frequent purpose of kiosks in secondary care settings was health information, followed by screening and patient registration. In primary care settings, the most frequent purpose was likewise health information, followed by clinical measurements and medication adherence. This agrees with the findings of a review by Joshi and Trout [42], where most (58%) health information kiosks were found in clinical settings. The review concluded that health information kiosks were feasible mediums for disseminating health information among various users in clinical and community settings, particularly if computer-based tailoring is used.

Table 2. Purposes of the most frequent settings (N=134).

Purpose	Settings, n (%)					Total, n (%)
	Secondary care (n=63)	Community (n=32)	Primary care (n=24)	Pharmacy (n=8)	Multiple (n=7)	
Health information	23 (37)	7 (22)	12 (50)	1 (13)	4 (57)	47 (35.1)
Clinical measurements	3 (5)	12 (28)	7 (29)	5 (63)	1 (14)	28 (20.9)
Screening	12 (19)	4 (13)	1 (4)	0 (0)	0 (0)	17 (12.7)
Telehealth	0 (0)	8 (25)	0 (0)	1 (13)	2 (29)	11 (8.2)
Patient registration	7 (11)	0 (0)	1 (4)	0 (0)	0 (0)	8 (6)
Patient feedback	5 (8)	0 (0)	1 (4)	0 (0)	0 (0)	6 (4.5)
Patient outcomes data	6 (10)	0 (0)	0 (0)	0 (0)	0 (0)	6 (4.5)
Medication reconciliation	3 (5)	0 (0)	2 (8)	0 (0)	0 (0)	5 (3)
Other	1 (2)	1 (3)	0 (0)	1 (13)	0 (0)	3 (2.2)
Patient triage	3 (5)	0 (0)	0 (0)	0 (0)	0 (0)	3 (2.2)

For kiosks installed in community settings and retail pharmacies, the most frequent purpose was to collect clinical measurements.

For kiosks installed in specialty and outpatient clinics in secondary care, sexual health was the most frequent condition addressed by kiosks (4/134, 3%) [50-53], followed by breastfeeding (3/134, 2.2%) [54,55], cancer (2/134, 1.5%) [56,57], chronic kidney disease (2/134, 1.5%) [58,59], HIV (2/134, 1.5%) [60,61], mental health (2/134, 1.5%) [62,63], and orthopedics (2/134, 1.5%) [64,65], with other conditions making up the remaining implementations (12/134, 9%), as shown in Table 3.

In kiosks deployed in EDs, the most frequently encountered health domains were HIV, acute care, and asthma. The HIV screening process in the ED was streamlined using kiosks (7/134, 5.2%) [66,67]. The privacy and relative anonymity of HIV screening via kiosks are reasons cited for the successful deployment of kiosks for this purpose, as patients preferred screening via kiosks rather than by a person, possibly as they felt more secure disclosing intimate details to a computer screen than to a person [68,69]. Kiosks were also able to increase patient knowledge about HIV testing [60,61,70]. Other conditions that were screened using kiosks were dementia (3/134, 2.2%), mental health (2/134, 1.5%), domestic violence/home safety (2/134, 1.5%), alcohol and drug use (1/134, 0.7%), dermatology (1/134, 0.7%), and urinary tract infection (1/134, 0.7%). Dementia screening took place in kiosks deployed in community settings, as well as for dermatology, in which the kiosk was equipped to take images of skin lesions [71]. One of the mental health screening kiosks was set in primary care and the other in secondary care, and all other screening kiosks were deployed in secondary care, mostly in

acute care/EDs. In the case of kiosks deployed in the community for screening, the situation is quite similar to asynchronous internet-based medical consultations.

Kiosks aided in the provision of acute care in the ED by performing patient triage, reliably collecting patient data, and significantly improving the time to identify new arrivals [72,73]. Other uses in the acute care pathway in the ED included patient registration [74] and medication adherence [75].

Primary care kiosks most frequently dealt with multiple conditions (7/134, 5.2%) [76-82], followed by cardiovascular disease (2/134, 1.5%) [83,84], general health and well-being (2/134, 1.5%) [85,86], hypertension (2/134, 1.5%) [87,88], and pediatric injuries (2/134, 1.5%) [89,90]. The community kiosks were for multiple conditions (8/134, 6%) and general health and well-being (2/134, 1.5%). Other conditions such as cardiovascular disease, dental health, dermatology, hypertension, infant mortality, pediatrics, and increasing social contact made up the rest (7/134, 5.2%). In studies where kiosks were deployed in pharmacies, the targeted health domains were hypertension [91-93], general health and well-being [94], and obesity [95]. In one of the studies, users accessed their personal health records through a kiosk at the pharmacy [96].

We examined the papers to determine if multiple papers evaluated the same kiosk system. A careful examination of the papers by authorship and system description revealed that 20.9% (28/134) of the primary studies were about 9 distinct kiosk systems. The 28 papers covered the settings, purposes, and conditions described in Table 4.

Thus, there were 115 distinct kiosk systems described in the 134 papers.

Table 3. Conditions and settings (N=134).

Condition	Setting, n (%)					Total, n (%)
	Secondary care (n=63)	Community (n=32)	Primary care (n=24)	Pharmacy (n=8)	Multiple (n=7)	
Multiple conditions	5 (8)	16 (50)	7 (29)	2 (25)	3 (43)	33 (24.6)
HIV	10 (16)	0 (0)	0 (0)	0 (0)	0 (0)	10 (7.5)
Hypertension	1 (2)	3 (9)	2 (8)	4 (50)	0 (0)	10 (7.5)
Pediatric injuries	4 (6)	1 (3)	2 (8)	0 (0)	0 (0)	7 (5.2)
Health and well-being	0 (0)	3 (9)	2 (8)	1 (13)	0 (0)	6 (4.5)
Medication	4 (6)	0 (0)	2 (8)	0 (0)	0 (0)	6 (4.5)
Cardiovascular disease	1 (2)	1 (3)	2 (8)	0 (0)	1 (14)	5 (3.7)
Mental health	3 (5)	0 (0)	1 (4)	0 (0)	0 (0)	4 (3)
Sexual health	4 (6)	0 (0)	0 (0)	0 (0)	0 (0)	4 (3)
Acute care—ED[a]	3 (5)	0 (0)	0 (0)	0 (0)	0 (0)	3 (2.2)
Breastfeeding	3 (5)	0 (0)	0 (0)	0 (0)	0 (0)	3 (2.2)
Cancer	2 (3)	0 ()	1 (4)	0 (0)	0 (0)	3 (2.2)
Dementia	0 (0)	3 (9)	0 (0)	0 (0)	0 (0)	3 (2.2)
Pediatrics	2 (3)	1 (3)	0 (0)	0 (0)	0 (0)	3 (2.2)
Smoking	1 (2)	0 (0)	1 (4)	0 (0)	1 (14)	3 (2.2)
Asthma	2 (3)	0 (0)	0 (0)	0 (0)	0 (0)	2 (1.5)
Chronic kidney disease	2 (3)	0 (0)	0 (0)	0 (0)	0 (0)	2 (1.5)
Diabetes	0 (0)	0 (0)	0 (0)	0 (0)	2 (29)	2 (1.5)
Domestic violence or home safety	2 (3)	0 (0)	0 (0)	0 (0)	0 (0)	2 (1.5)
Obesity	0 (0)	0 (0)	1 (4)	1 (13)	0 (0)	2 (1.5)
Orthopedics	2 (3)	0 (0)	0 (0)	0 (0)	0 (0)	2 (1.5)
Alcohol and drug use	1 (2)	0 (0)	0 (0)	0 (0)	0 (0)	1 (0.7)
Cervical cancer	0 (0)	0 (0)	1 (4)	0 (0)	0 (0)	1 (0.7)
Childhood obesity	0 (0)	0 (0)	1 (4)	0 (0)	0 (0)	1 (0.7)
Dental health	0 (0)	1 (3)	0 (0)	0 (0)	0 (0)	1 (0.7)
Dermatology	0 (0)	1 (3)	0 (0)	0 (0)	0 (0)	1 (0.7)
Dog bites	1 (2)	0 (0)	0 (0)	0 (0)	0 (0)	1 (0.7)
Environmental health	1 (2)	0 (0)	0 (0)	0 (0)	0 (0)	1 (0.7)
Food safety	1 (2)	0 (0)	0 (0)	0 (0)	0 (0)	1 (0.7)
General medicine	1 (2)	0 (0)	0 (0)	0 (0)	0 (0)	1 (0.7)
Genetic study	1 (2)	0 (0)	0 (0)	0 (0)	0 (0)	1 (0.7)
Health care environment	1 (2)	0 (0)	0 (0)	0 (0)	0 (0)	1 (0.7)
Infant mortality	0 (0)	1 (3)	0 (0)	0 (0)	0 (0)	1 (0.7)
Organ donation	0 (0)	0 (0)	1 (4)	0 (0)	0 (0)	1 (0.7)
Patient communication	1 (2)	0 (0)	0 (0)	0 (0)	0 (0)	1 (0.7)
Radiology	1 (2)	0 (0)	0 (0)	0 (0)	0 (0)	1 (0.7)
Rehabilitation	1 (2)	0 (0)	0 (0)	0 (0)	0 (0)	1 (0.7)
Social contact	0 (0)	1 (3)	0 (0)	0 (0)	0 (0)	1 (0.7)
UTI[b]	1 (2)	0 (0)	0 (0)	0 (0)	0 (0)	1 (0.7)

[a]ED: emergency department.

XSL•FO
RenderX

[b]UTI: urinary tract infection.

Table 4. Kiosk systems described by multiple papers (N=28).

Kiosk system name	Country	Papers, n (%)	Settings	Purposes	Conditions
Telehealth Wellness Kiosk [97,98]	Unites States	2 (7)	Community	Telehealth	Multiple
HPV Project Kiosk [52,53]	Unites States	2 (7)	Secondary care	Health information	Sexual health
HIV Screening Kiosk [66-69,99-102]	Unites States	8 (29)	Secondary care	Screening	HIV
APHID Kiosk [103-106]	Unites States	4 (14)	Primary or secondary care	Medication adherence	Medication
PEMT Kiosk [54,55,107]	Unites States	3 (11)	Secondary care	Health information	Breastfeeding
My Kidney Care Centre [58,59]	Canada	2 (7)	Secondary care	Patient outcomes	Chronic kidney disease
KIO kiosk [108,109]	Unites States	3 (11)	Community	Screening or patient outcomes	Dementia
e-KISS kiosk [50,51]	Unites States	2 (7)	Secondary care	Health information	Sexual health
Safety in Seconds kiosk [110,111]	Unites States	2 (7)	Secondary care	Health information	Pediatric injuries

Country of Kiosk Installation

The countries where the 115 kiosk systems were deployed and their corresponding settings are listed in Table 5.

Table 5. Countries and settings of included studies (N=134).

Country	Community (n=32), n (%)	Multiple (n=7), n (%)	Pharmacy (n=8), n (%)	Primary care (n=24), n (%)	Secondary care (n=63), n (%)	Total, n (%)
United States[a]	17 (53)	6 (86)	3 (38)	15 (63)	40 (63)	81 (60.4)
Canada[a]	0 (0)	0 (0)	2 (25)	0 (0)	3 (5)	5 (3.7)
United Kingdom[a]	2 (6)	0 (0)	1 (13)	2 (8)	1 (2)	6 (4.5)
Germany[a]	0 (0)	0 (0)	1 (13)	0 (0)	2 (3)	3 (2.2)
India[b]	2 (6)	0 (0)	0 (0)	0 (0)	1 (2)	3 (2.2)
South Korea[a]	1 (3)	0 (0)	0 (0)	0 (0)	2 (3)	3 (2.2)
New Zealand[a]	2 (6)	0 (0)	0 (0)	0 (0)	1 (2)	3 (2.2)
Portugal[a]	0 (0)	1 (14)	0 (0)	1 (4)	0 (0)	2 (1.5)
Singapore[a]	0 (0)	0 (0)	0 (0)	2 (8)	0 (0)	2 (1.5)
Australia[a]	1 (3)	0 (0)	0 (0)	0 (0)	0 (0)	1 (0.7)
Brazil[c] and Portugal[a]	1 (3)	0 (0)	0 (0)	0 (0)	0 (0)	1 (0.7)
Japan[a]	1 (3)	0 (0)	0 (0)	0 (0)	0 (0)	1 (0.7)
Kenya[b]	1 (3)	0 (0)	0 (0)	0 (0)	0 (0)	1 (0.7)
The Philippines[b]	0 (0)	0 (0)	1 (13)	0 (0)	0 (0)	1 (0.7)
Sweden[a]	0 (0)	0 (0)	0 (0)	1 (4)	0 (0)	1 (0.7)
United States and Canada[a]	1 (3)	0 (0)	0 (0)	0 (0)	0 (0)	1 (0.7)

[a]High-income country.

[b]Lower middle–income country.

[c]Upper middle–income country.

Most kiosk studies were conducted in the United States, accounting for 70.4% (81/115) of the installed kiosk systems.

Of the 115 installed kiosk systems, Canada and the United Kingdom had 5 (4.3%) and 6 (5.2%) systems, respectively, and

Germany, India, South Korea, and New Zealand contributed 3 (2.6%) systems each. The list includes 11 high-income countries (Australia, Canada, Germany, Japan, New Zealand, Portugal, Singapore, South Korea, Sweden, the United Kingdom, and the United States), 1 upper middle–income country (Brazil), and 3 lower middle–income countries (India, Kenya, and the Philippines), as classified by the World Bank [112]. On the basis of the included primary studies, high-income countries had a higher proportion of kiosks situated in secondary care, whereas upper and lower middle–income countries had a greater proportion of kiosks deployed in primary care, the community, and pharmacies.

Number of Studies Published Per Year

The studies included in the review were published in the period covering 2009 to 2020 (Table 6). The included 134 primary studies represent an almost 6-fold increase from the 25 studies cited by the review by Jones [1] published in 2009.

Table 6. Number of primary studies published per year from 2009 to 2020 (N=134).

Year	Studies, n (%)
2009	5 (3.7)
2010	10 (7.5)
2011	13 (9.7)
2012	8 (6)
2013	19 (14.2)
2014	17 (12.7)
2015	13 (9.7)
2016	11 (8.2)
2017	14 (10.4)
2018	12 (9)
2019	8 (6)
2020	4 (3)

Type of Kiosk Access

Most (94/134, 70.1%) of the kiosks described in the included papers were integrated into clinical pathways (Table 7) and were cited mostly in secondary care (specialty clinics, EDs, and hospital inpatient clinics) and primary care facilities. The most common uses of these kiosks were delivering health information, clinical measurements, and screening. Opportunistic kiosks were described in approximately a quarter of the included studies and were most often found in community settings, clinical settings, and pharmacies. The most frequent uses of opportunistic kiosks were for delivering health information and taking clinical measurements.

Table 7. Type of access to health kiosk (N=134).

Setting	Type of access, n (%)			Total, n (%)
	Integrated (n=94)	Opportunistic (n=32)	Both (n=8)	
Secondary care	54 (57)	9 (28)	0 (0)	63 (47)
Community	16 (17)	11 (34)	5 (63)	32 (23.9)
Primary care	18 (19)	5 (16)	1 (13)	24 (17.9)
Pharmacy	3 (3)	5 (16)	0 (0)	8 (6)
Multiple	3 (3)	2 (6)	2 (25)	7 (5.2)

Patient Self–check-in Kiosks

One type of kiosk that has been widely deployed over the past decade is the patient self–check-in kiosk in general practices, outpatient clinics, and hospitals. In the United Kingdom, the rise of the electronic patient self–check-in kiosk can be traced to a guide released by the National Health Service (NHS) in 2009, entitled *Improving access, responding to patients: A "how-to" guide for GP practices.* The guide included a section on *self-service check-in screens*, which would allow patients to check themselves in for an appointment quickly [113]. The guide included practical tips on deployment, including estimated acquisition and maintenance costs. We could not find any official figures for the number of self–check-in kiosks in general practices and hospitals in the United Kingdom. The best data we could find was that a vendor of patient self-check-in software for GP surgeries estimated that their system had been used 30 million times since April 2018 [114]. Given that there are approximately 300 million GP appointments per year in the NHS [115], this vendor would account for 5% of patient appointments in the NHS in a 2-year period.

We found only a few studies in our literature search that evaluated patient self–check-in kiosks. These studies showed statistically significant reductions in waiting times for patients who checked in using the kiosks compared with those who did not [65,74]. What was surprising was the small number of studies in the published academic literature, given the growing adoption of patient self–check-in screens over the past 10 years. However, it may be that the studies were performed as service evaluations rather than academic research and not submitted for academic publication.

Reporting on the Usability Evaluation of Kiosks

Of the 7 reviews retrieved, 3 (43%) mentioned usability as one of the outcomes reported in their included studies [40,42,43]. However, none of the reviews in the included literature mentioned the types of usability evaluation methods used in the included studies. Usability evaluations of health kiosks were reported in 20.1% (27/134) of the included primary studies, slightly higher than the 16% reported in a systematic review by Joshi and Trout [42].

The methods used for usability evaluation in 27 studies are listed in Table 8.

Table 8. Usability evaluations of health kiosks (N=27).

Methods	Values, n (%)
Questionnaires	13 (48)
Validated questionnaires	3 (11)
Focus groups	3 (11)
Interviews	7 (26)
Completion rates	4 (15)
Error rates	2 (7)
Multiple methods	10 (37)
Heuristic evaluation	3 (11)
Think-aloud	8 (22)
Click recording	2 (7)
Visual observation	5 (19)

Although questionnaires were the most frequently used usability evaluation method, only 11% (3/27) of studies used validated questionnaires, namely the System Usability Scale, the Technology Acceptance Model, and the Perceived Usefulness/Perceived Ease of Use questionnaire. Validated questionnaires enable researchers to compare their results with those of other studies. Questionnaires are subjective and quantitative methods. Some of the studies used qualitative methods such as focus groups, interviews, behavioral observations, and think-aloud sessions (8/27, 22%). Qualitative methods are usually used during the developmental stages. Objective methods were also used, such as completion times, error rates, and click recordings. Heuristic evaluation, using a checklist of desired heuristic features, was used only in a small minority of the studies (3/27, 11%). Approximately half of the studies (10/27, 37%) used >1 method of usability evaluation. Approximately 37% (10/27) of usability evaluations of health kiosks were able to identify usability issues. Most (16/27, 59%) reported that the users found the health kiosks easy to use.

Telemonitoring Kiosks

Approximately 7.5% (10/134) of papers described the use of kiosks to deliver some form of telemonitoring or teleconsultations between 2011 and 2014. Most papers (6/10, 60%) described kiosks implemented in retirement communities for the use of older adults. Approximately 30% (3/10) of papers related to the same kiosk for a residential community of older adults in New Zealand [97,98,116]. Another kiosk was implemented in an urgent care pharmacy [27], and another, aimed at community-dwelling older adults, was tested in a laboratory setting [117]. One of the kiosks, not yet widely implemented, was deployed in a rural community health center [118].

Approximately 80% (8/10) of papers described 5 different kiosks that provided telemonitoring services, including monitoring of vital signs such as BP and oximetry. These kiosks included a screen but did not allow for 2-way live communication with a health care provider. Telemonitoring kiosks aimed at older adults often also included measures of cognitive performance and the opportunity for residents to engage with educational videos and *brain fitness* games. Health information collected by the kiosk was transmitted electronically to relevant health care professionals who could monitor ongoing conditions such as hypertension [119] and cognitive decline [120,121]. In some cases, users were also able to download their information and observe changes over time [97,98,116]. A kiosk designed for a rural community center in India, although not yet widely implemented, also included functions that enable the detection of malaria and tuberculosis and upload of radiology images [118].

Teleconsultation Kiosks

Of the 20% (2/10) of papers that outlined kiosks that offered the opportunity for users to interact in a live 2-way consultation with a health care professional, one of them, HealthSpot, is no longer in operation. We will discuss the history of HealthSpot

in greater detail in the following sections. The kiosk that is still in operation has been implemented in 7 urgent care pharmacies across New York City and included audiovisual equipment enabling a web-based consultation with an ED physician, a BP cuff, a pulse oximeter, and a thermometer [27]. This provider also offered the same service but via a mobile app. The authors reported that out of a total of 1996 web-based consultations conducted, only 238 were at kiosks, and the daily use of each kiosk location was low. However, people who used the kiosks were less likely to experience technical difficulties compared with those who used the app. Interestingly, the authors also found that those who used the kiosks were significantly more likely to be visitors to the area than local people, suggesting that a visit to the kiosk represented an opportunity to access care when not familiar with local services.

Training Needs for Implementing Kiosks for Telemonitoring and Teleconsultations

Only 20% (2/10) of papers detailing kiosks providing telemonitoring or teleconsultation services described the training required to implement the kiosk. Wilamowska et al [116] briefly noted that the kiosk vendor organized 2 training sessions to familiarize the research team members with the design and details of the kiosk and its output data. Training for end users (older adults) was not described [116].

Resnick et al [119] described how their kiosk for older adults incorporated training for both researchers and end users [119]. Retirement center employees and researchers were first taught how to use the device by the kiosk developers. The research staff then trained older people on how to use the kiosk equipment. No further details on what the training involved were included in the paper. However, nearly all older adults reported being *very comfortable* with the technology; 81% reported that it was easy to use, and 98% reported that they would recommend it to others. However, analysis of compliance data revealed that kiosk use decreased over time, and the authors suggested that enhanced training on the use of equipment may facilitate the continued use of the kiosk following the initial *honeymoon* period.

Barriers to and Enablers of Teleconsultation Kiosk Adoption

The experience of the telemedicine kiosk pioneer HealthSpot provides a good understanding of barriers to adoption. HealthSpot was founded in 2010 and raised approximately US $46.7 million in funding. It also attracted several big-name partners such as Xerox, MetroHealth, Mayo Health, Kaiser Permanente, the Cleveland Clinic, and Rite Aid (the third largest retail pharmacy chain in the United States). HealthSpot's telemedicine kiosk was fully enclosed and used proprietary cloud-based software and was equipped with high-definition videoconferencing, a BP cuff, thermometer, stethoscope, otoscope, dermatoscope, and a built-in weighing scale [122]. Despite this promising start, HealthSpot ceased operations in December 2015.

Mudumba [123] and Chen [7] enumerated the following reasons for the failure of HealthSpot:

- Too much time spent on the academic validation of kiosk functionality rather than vetting the business model in the market
- Requires prescheduling appointments for HealthSpot kiosk users, which goes against the utility aspect of telehealth
- No integration with mobile health platforms
- Inadequate planning for scaling
- The target market was too small

The HealthSpot kiosk used proprietary videoconferencing software, whose high cost weakened the HealthSpot business model. According to Chen [7], kiosks need to cost <US $5000 per unit for the business model to succeed. These lessons must be considered when companies attempt to enter the telehealth kiosk market.

Some other studies also mentioned barriers to and enablers of kiosk adoption. Venkatesh [23] noted that advice from strong and weak ties was an enabler of kiosk adoption by mothers. Conversely, hindrance from strong and weak ties was a barrier to kiosk adoption [23]. Ackerman [124] investigated the reasons for nonadoption of a kiosk to screen for urinary tract infection in an ED setting. The kiosk had previously been successfully adopted in an urgent care clinic setting. The research showed that kiosk algorithms were not adaptable to changing situations in a busy emergency room. The researchers also failed to involve triage nurses in the development of the system, which resulted in disengagement and a nonsupportive attitude toward the kiosk.

Discussion

Principal Findings

In this review, we sought to describe the current roles that health kiosks play in the health care system in terms of settings, purposes, health domains, and type of kiosk (opportunistic or integrated into a care pathway), as reported in the existing literature. We also investigated the use of kiosks for patient self–check-in, the extent of reporting of the usability evaluation of health kiosks, and the factors that affect the use of kiosks for remote consultations. We identified that clinical settings still comprised most (87/134, 64.9%) sites for health kiosks, and community settings accounted for some (32/134, 23.9%) of the kiosk installations in the included studies. Retail pharmacy settings comprised 5.9% (8/134) of the included studies. However, BP kiosks have long been deployed in pharmacies for quite some time. In 2012, Alpert [91] reported that 1 million BP readings per day were recorded in BP kiosks in pharmacies. Currently, kiosks with more functions are being deployed in pharmacies, including drug dispensing [125], teleconsultation [13], drug disposal, and health measurements and information kiosks [126].

Comments on the Findings

Country of Kiosk Installation, Clinical Integration, Increase in Publication, and Usability of Kiosks

When looking at the countries of installation, high-income countries dominate in the studies on health kiosks included in our review, accounting for 73% (11/15) of the countries where kiosks were installed. Regarding countries and settings, it can

be noted that high-income countries have a larger proportion of kiosks in secondary care settings, whereas upper and lower middle–income countries tend to have their kiosks installed in community and primary care settings. This reflects the more advanced health infrastructure of high-income countries, which can afford to deploy information technology solutions in their health systems. A study on barriers to and facilitators of the deployment of health kiosks in Iran, an upper middle–income country, listed a lack of resources as one of the barriers [127]. A report from the World Health Organization / World Bank in 2017 stated that half of the world's population still lacks access to essential health services [128]. In situations where health resources are in short supply, kiosks will probably not be high in the list of priorities.

Kiosks are more likely to be integrated into a clinical pathway (94/134, 70.1%), especially if they were in a clinical setting. Community kiosk installations were evenly divided between integrated access and opportunistic/dual access. In both clinical and community settings, health information and clinical measurements were the most frequent purposes for kiosks.

The 6-fold increase in publications on health kiosks is an indication of the growing use of computerized kiosks in health care. This also coincides with the increased growth of the computerized kiosk market in other sectors, such as retail, hospitality, and banking, during the same period [129]. However, there has been a drop in the number of publications per year since 2013, which could lead to the conclusion that there has been a decrease in the relevance and interest in health kiosks since that year. However, another explanation could be that because of the continued growth of the use of health kiosks and self-service kiosks in general since 2010, as stated in market research reports, the use of health kiosks has become more normalized since 2013, such that fewer researchers are publishing work in this area in the same way that there are few research papers about airline check-in kiosks and automated teller machines.

The proportion of health kiosk studies that include a usability evaluation of the kiosk has not changed much since 2014 and is in the minority (<20%). This is consistent with the low rate of reporting on usability evaluations of digital health technologies in general [31]. There is also a lack of use of validated questionnaires for usability evaluations, making comparisons of usability between studies difficult. As user experience evaluations are now required for the commissioning of new digital health devices [130], manufacturers who wish to enter and develop products in the growing health kiosk market will need guidance, capacity, and capability building in user experience evaluation.

Limitations and Strengths of the Review

This review has a few limitations. We were only able to search for papers published in English, which may have excluded several papers about health kiosks that were not published in English. This means that we were not able to include papers about kiosks installed in countries such as China, Japan, South Korea, and others if they were published in a language other than English. We were also constrained to reduce our search terms, as the use of the term *touchscreen* (as was done in the

2009 review by one of the authors) resulted in the inclusion of many papers on smartphones and tablets, which were clearly not kiosks. In addition, the term *kiosk* is not part of a controlled vocabulary (eg, Medical Subject Heading). We deliberately excluded papers on proposed kiosks, including only papers on actual kiosk installations. Some of these kiosk proposals may have become actual kiosks in the interim; however, we would have no way of knowing which one was successfully implemented. The quick pace of technological change also outstrips the pace of academic publishing; hence, we also included information gathered from web search engines and key informants. Finally, the competitive nature of digital health technology makes information about development methods closely guarded trade secrets, which makes the publication of these methods in academic journals unlikely.

We were aware that there were existing reviews on health kiosks before we started this scoping review. Our search identified 7 prior reviews, the latest of which was published in 2013. It was our consensus that in the 7 years since the last review, there were sufficient technological advances to warrant a new review. In the process of writing the findings of this review, a new systematic review of integrated health kiosks was published [131]. This review only covered publications up to 2018 and only included 37 articles. Our review covers publications from January 2009 to June 2020 and includes 137 articles. Thus, one of the strengths of our review is that it is more timely and comprehensive and complements the findings of previously published reviews.

Implications of the Findings

Kiosk Adoption: Barriers and Enablers and Training Needs

A recent qualitative study of 20 experts in Iran investigated their perceptions of the barriers to and facilitators of health kiosk adoption [127]. They identified lack of resources, low digital literacy, and resistance from health system officials as some of the barriers to adoption. On the other hand, high internet and electric power penetration rates, deployment of telemedicine, and integrated management of health services were cited as facilitators for adoption. The barriers to and enablers of kiosk adoption were mentioned in only a few of the studies included in our review; thus, there is a need for further research on this topic.

The current success of MedicSpot in the United Kingdom contrasts greatly with the failure of HealthSpot in the United States. MedicSpot follows the points made by Chen [7] and Mudumba [123] by allowing walk-in consultations, integrating with a mobile platform, and planning carefully for scaling. MedicSpot was shortlisted for the Digital Innovation Team of the Year at the 2019 British Medical Journal Awards [16].

This brings us to the need for training in using health kiosks for teleconsultation. Although most of the included papers about kiosks for telemonitoring and teleconsultations were aimed at older adults with less technical experience, it is surprising that end user training needs are not frequently described in more detail. It is possible that kiosk use with touch screens has been normalized in other areas of daily living (eg, banking and supermarket shopping), and thus, their use is seen to be intuitive.

The lack of training may also reflect that, in some cases, a kiosk may be accompanied by a trained *person* to support the use and management of technical issues. This may be in accordance with previous NHS guidance that recent technology be introduced together with someone who can assist inexperienced users [113]. This is also being practiced in the Danish *chronic obstructive pulmonary disease briefcase* telemedicine intervention, where the patient's equipment was installed by a technician who also provided instructions on how to switch the system on and off and how to position the finger clip pulse oximeter [132]. The learning needs of health professionals in using video calls to support patients have been successfully identified through workshops [133]. A similar methodology can be used to create training programs for health care professionals to use video calls for teleconsultations.

Patient Self–check-In Kiosks

The adoption of patient self–check-in kiosks has had its share of criticism and negative news reports. An opinion piece by Williamson [134] warned that the impersonality of these systems is contrary to general practice's emphasis on personal and therapeutic relationships. Most practices have responded to this by still giving patients the option of checking in for their appointments via a human receptionist. There have also been concerns about the display of personal information on kiosk screens that could be viewed by others [135], as well as the hygiene implications of multiple users touching the same kiosk. The solutions to this are limiting the display of information to the appointment time, health care provider, and examining room and by providing hand sanitizing gel and regularly disinfecting the kiosk screen. Further research on no-touch interfaces with kiosks, such as voice and gestures, can also decrease the possibility of spreading infections [136,137]. Another news item in 2016 reported that some patients exaggerated their symptoms when answering questions at self–check-in kiosks

installed in the accident and ED of a hospital to jump the queue. The hospital responded by combining the use of the electronic system with face-to-face input from senior clinicians to ensure that more accurate information was gathered [138]. This points to the need for a regular audit of the security and privacy of the health kiosk installation. Some research has been conducted on ensuring the security and privacy of kiosks [139]; however, more work will be needed as the number of health kiosk installations increases. Security is related to the need to regularly update kiosk software to respond to security threats, as well as to meet changing needs as health care situations evolve. A cloud-based platform may be a solution; however, it also creates the need for a constant connection to the internet. All these issues require further research.

Conclusions

In conclusion, this review characterizes the present roles that health kiosks play in the health care system based on the existing literature. We have established that despite the growth in erstwhile health kiosk replacements such as personal smart devices and their attendant apps, health kiosks still have a vital role to play in the health care system, such as in the collection of clinical measurements for teleconsultations, provision of access to eHealth for the older population without smartphones, and provision of tailored and vetted health information at the point of service. We also identified research gaps such as identifying training needs for using the kiosk/video call combination for teleconsultations; methods for usability testing of kiosks; barriers to and enablers of kiosk deployment; and the exact extent of kiosk use for patient self–check-in for primary, secondary, and tertiary care. We also recommend the implementation of programs that will increase the capability and capacity of kiosk developers to perform user experience evaluations, both during development and while in service.

Authors' Contributions

RJ conceived of the study. IDM, DA, and KE screened the studies and conducted the data extraction. IDM analyzed the extracted data. The review was written by IDM, DA, and KE, with revisions from RJ, AC, and EM. RJ, AC, and EM provided feedback on the draft text. The authors confirmed that they followed all appropriate research reporting guidelines. The checklist for scoping reviews was uploaded in Multimedia Appendix 2.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Summary of included studies.
[PDF File (Adobe PDF File), 526 KB - medinform_v10i3e26511_app1.pdf]

Multimedia Appendix 2
PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) checklist.

[PDF File (Adobe PDF File), 150 KB - medinform_v10i3e26511_app2.pdf]

References

1. Jones R. The role of health kiosks in 2009: literature and informant review. Int J Environ Res Public Health 2009;6(6):1818-1855 [FREE Full text] [doi: 10.3390/ijerph6061818] [Medline: 19578463]
2. Internet access – households and individuals, Great Britain: 2020. Office for National Statistics. 2020. URL: https://www.ons.gov.uk/peoplepopulationandcommunity/householdcharacteristics/homeinternetandsocialmediausage/bulletins/internetaccesshouseholdsandindividuals/2020 [accessed 2020-09-07]
3. Internet users, UK: 2019. Office for National Statistics. 2019. URL: https://www.ons.gov.uk/businessindustryandtrade/itandinternetindustry/bulletins/internetusers/2019#generation-gap-narrowing-in-recent-internet-use [accessed 2020-09-07]
4. UK communications market report. Ofcom. 2018. URL: https://www.ofcom.org.uk/__data/assets/pdf_file/0022/117256/CMR-2018-narrative-report.pdf [accessed 2022-03-23]
5. Franklin R. 11 surprising mobile health statistics. Mobius MD. 2021. URL: https://www.mobius.md/blog/2019/03/11-mobile-health-statistics/ [accessed 2020-09-09]
6. Statistics. International Telecommunication Union. URL: https://www.itu.int/en/ITU-D/Statistics/Pages/stat/default.aspx [accessed 2021-02-26]
7. Chen M. How to make money in telehealth. Telehealth Med Today 2018;1(3):-. [doi: 10.30953/tmt.v1.83]
8. Kochelek K. Why healthcare kiosks are essential in the changing medical landscape. Frank Mayer. 2020. URL: https://www.frankmayer.com/blog/why-healthcare-kiosks-are-essential-in-the-changing-medical-landscape/ [accessed 2020-10-22]
9. Boyle M. Mobile internet statistics. Finder. 2021. URL: https://www.finder.com/uk/mobile-internet-statistics [accessed 2020-09-14]
10. A decade of digital dependency. Ofcom. 2018. URL: https://www.ofcom.org.uk/about-ofcom/latest/features-and-news/decade-of-digital-dependency [accessed 2020-09-15]
11. Lynch P, Wainwright D. Coronavirus: how GPs have stopped seeing most patients in person. BBC News. 2020. URL: https://www.bbc.co.uk/news/uk-england-52216222 [accessed 2020-09-15]
12. Downey A. Text and telephone consultations trump video during COVID-19. Digital Health. URL: https://www.digitalhealth.net/2020/06/text-and-telephone-consultations-trump-video-during-covid-19/ [accessed 2020-09-15]
13. Private COVID-19 testing and GP services. Medicspot. URL: https://www.medicspot.co.uk/ [accessed 2021-02-25]
14. MedicSpot Virtual. NHS for Sale. URL: https://www.nhsforsale.info/private-providers/medicspot-virtual-new/ [accessed 2021-02-25]
15. Mageit S. Asda launches virtual in-store GP service with Medicspot. Healthcare IT News. 2020. URL: https://www.healthcareitnews.com/news/emea/asda-launches-virtual-store-gp-service-medicspot [accessed 2021-02-25]
16. Wise J. BMJ awards 2019: digital innovation team of the year. BMJ 2019;365:l1519. [doi: 10.1136/bmj.l1519] [Medline: 30944089]
17. Telemedicine equipment. Amwell. URL: https://business.amwell.com/telemedicine-equipment/kiosks/ [accessed 2020-07-24]
18. Lovett L. Amwell scores $194M, as telehealth business booms during coronavirus pandemic. MobiHealthNews. 2020. URL: https://www.mobihealthnews.com/news/amwell-scores-194m-telehealth-business-booms-during-coronavirus-pandemic [accessed 2020-08-24]
19. Nittas V. When eHealth goes viral: the strengths and weaknesses of health tech during COVID-19. MobiHealthNew. 2020. URL: https://www.mobihealthnews.com/news/europe/when-ehealth-goes-viral-strengths-and-weaknesses-health-tech-during-covid-19 [accessed 2020-09-15]
20. Covid 19 reception kiosk. Elephant Kiosk. URL: https://web.archive.org/web/20210514154151/http://www.elephantkiosks.co.uk/covid_reception_kiosk/ [accessed 2020-08-27]
21. The first connected local telemedicine booth. The Consult Station. URL: https://www.h4d.com/en-uk/connected-telemedicine-booth/ [accessed 2020-07-24]
22. Mageit S. H4D raises €15m to improve access to healthcare with its telemedicine pod, the Consult Station. MobiHealthNews. 2020. URL: https://www.mobihealthnews.com/news/europe/h4d-raises-15m-improve-access-healthcare-its-telemedicine-pod-consult-station [accessed 2020-07-24]
23. Venkatesh V, Rai A, Sykes TA, Aljafari R. Combating infant mortality in rural India: evidence from a field study of eHealth kiosk implementations. MIS Q 2016;40(2):353-380. [doi: 10.25300/misq/2016/40.2.04]
24. Gagnon MP, Fortin JP, Landry R. Telehealth to support practice in remote regions: a survey among medical residents. Telemed J E Health 2005;11(4):442-450. [doi: 10.1089/tmj.2005.11.442] [Medline: 16149890]
25. Palozzi G, Schettini I, Chirico A. Enhancing the sustainable goal of access to healthcare: findings from a literature review on telemedicine employment in rural areas. Sustainability 2020;12(8):3318. [doi: 10.3390/su12083318]
26. Wherton J, Greenhalgh T. Evaluation of the attend anywhere / near me video consulting service in Scotland, 2019-20. Scottish Government. 2020. URL: http://www.gov.scot/publications/evaluation-attend-anywhere-near-video-consulting-service-scotland-2019-20-main-report/ [accessed 2021-02-25]

XSL·FO

RenderX

27. Nachum S, Gogia K, Clark S, Hsu H, Sharma R, Greenwald PW. An evaluation of kiosks for direct-to-consumer telemedicine using the national quality forum assessment framework. Telemed J E Health 2021;27(2):178-183. [doi: 10.1089/tmj.2019.0318] [Medline: 32589518]

28. Cornwall: supplying skills for the local visitor economy. Local Government Association. 2019. URL: https://local.gov.uk/cornwall-supplying-skills-local-visitor-economy [accessed 2020-09-14]

29. Living with tourists: The impact of tourism on health services in Cornwall: a report of the impact of tourism upon health care in Cornwall - single issue panel. Cornwall County Council. 2004. URL: https://web.archive.org/web/20151016140132/https://www.cornwall.gov.uk/media/3628063/4_e_impact_of_tourism_on_health_services_in_cornwall.pdf [accessed 2022-03-25]

30. Bastien JM. Usability testing: a review of some methodological and technical aspects of the method. Int J Med Inform 2010;79(4):e18-e23. [doi: 10.1016/j.ijmedinf.2008.12.004] [Medline: 19345139]

31. Maramba I, Chatterjee A, Newman C. Methods of usability testing in the development of eHealth applications: a scoping review. Int J Med Inform 2019;126:95-104. [doi: 10.1016/j.ijmedinf.2019.03.018] [Medline: 31029270]

32. Broekhuis M, van Velsen L, Hermens H. Assessing usability of eHealth technology: a comparison of usability benchmarking instruments. Int J Med Inform 2019;128:24-31. [doi: 10.1016/j.ijmedinf.2019.05.001] [Medline: 31160008]

33. Blignault I, Kennedy C. Training for telemedicine. J Telemed Telecare 1999;5 Suppl 1:S112-S114. [doi: 10.1258/1357633991932793] [Medline: 10534864]

34. Pham MT, Rajić A, Greig JD, Sargeant JM, Papadopoulos A, McEwen SA. A scoping review of scoping reviews: advancing the approach and enhancing the consistency. Res Synth Methods 2014;5(4):371-385 [FREE Full text] [doi: 10.1002/jrsm.1123] [Medline: 26052958]

35. Daudt HM, van Mossel C, Scott SJ. Enhancing the scoping study methodology: a large, inter-professional team's experience with Arksey and O'Malley's framework. BMC Med Res Methodol 2013;13:48 [FREE Full text] [doi: 10.1186/1471-2288-13-48] [Medline: 23522333]

36. Jones R, Asthana S, Walmsley A, Sheaff R, Andrade J, May J, et al. Developing the eHealth sector in Cornwall. University of Plymouth. 2019. URL: https://www.plymouth.ac.uk/uploads/production/document/path/14/14515/Developing_the_eHealth_sector_in_Cornwall.pdf [accessed 2022-03-23]

37. Kreuter MW, Black WJ, Friend L, Booker AC, Klump P, Bobra S, et al. Use of computer kiosks for breast cancer education in five community settings. Health Educ Behav 2006;33(5):625-642. [doi: 10.1177/1090198106290795] [Medline: 16923835]

38. R: a language and environment for statistical computing. R Core Team. 2020. URL: https://www.r-project.org/ [accessed 2022-03-23]

39. McIndoe R. Health kiosk technologies. In: Brown SA, Brown M, editors. Ethical issues and security monitoring trends in global healthcare: technological advancements. Hershey, PA: IGI Global; 2011.

40. Joshi A, Trout K. Heath kiosks as an equal opportunity resource for better health: a systematic review. In: The First International Conference on Global Health Challenges. 2012 Presented at: Global Health '12; October 21-26, 2012; Venice, Italy URL: https://pdfs.semanticscholar.org/26a2/63f9f71d446e2541ddbd413d671220e17fc9.pdf

41. Yvonne Chan YF, Nagurka R, Bentley S, Ordonez E, Sproule W. Medical utilization of kiosks in the delivery of patient education: a systematic review. Health Promot Perspect 2014;4(1):1-8 [FREE Full text] [doi: 10.5681/hpp.2014.001] [Medline: 25097831]

42. Joshi A, Trout K. The role of health information kiosks in diverse settings: a systematic review. Health Info Libr J 2014;31(4):254-273 [FREE Full text] [doi: 10.1111/hir.12081] [Medline: 25209260]

43. Nicholas D, Huntington P, Williams P. Digital health information for the consumer: evidence and policy implications. Milton Park, UK: Routledge; 2016.

44. Al Hamarneh YN, Houle SK, Chatterley P, Tsuyuki RT. The validity of blood pressure kiosk validation studies: a systematic review. Blood Press Monit 2013;18(3):167-172. [doi: 10.1097/MBP.0b013e328360fb85] [Medline: 23635486]

45. Abraham O, Patel M, Feathers A. Acceptability of health kiosks within African American community settings: a pilot study. Health Serv Res Manag Epidemiol 2018;5:2333392817752211 [FREE Full text] [doi: 10.1177/2333392817752211] [Medline: 29383325]

46. Dulchavsky SA, Ruffin WJ, Johnson DA, Cogan C, Joseph CL. Use of an interactive, faith-based kiosk by congregants of four predominantly, African-American churches in a metropolitan area. Front Public Health 2014;2:106 [FREE Full text] [doi: 10.3389/fpubh.2014.00106] [Medline: 25140296]

47. Bean K, Davis O, Valdez H. Bridging the digital divide: a bilingual interactive health kiosk for communities affected by health disparities. J Commun Inform 2013;9(2):1-7 [FREE Full text] [doi: 10.15353/joci.v9i2.3169]

48. Matthews PH, Darbisi C, Sandmann L, Galen R, Rubin D. Disseminating health information and diabetes care for Latinos via electronic information kiosks. J Immigr Minor Health 2009;11(6):520-526. [doi: 10.1007/s10903-008-9134-6] [Medline: 18392935]

49. Leeman-Castillo B, Beaty B, Raghunath S, Steiner J, Bull S. LUCHAR: using computer technology to battle heart disease among Latinos. Am J Public Health 2010;100(2):272-275 [FREE Full text] [doi: 10.2105/AJPH.2009.162115] [Medline: 20019305]

50. Shafii T, Benson SK, Morrison DM, Hughes JP, Golden MR, Holmes KK. Results from eKISS (electronic KIOSK Intervention for Safer-Sex): a pilot randomized controlled trial to test an interactive computer-based intervention for sexual health in adolescents and young adults. J Adolesc Health 2014;54(2):S10 [FREE Full text] [doi: 10.1016/j.jadohealth.2013.10.036]

51. Shafii T, Benson SK, Morrison DM, Hughes JP, Golden MR, Holmes KK. Results from e-KISS: electronic-KIOSK intervention for safer sex: a pilot randomized controlled trial of an interactive computer-based intervention for sexual health in adolescents and young adults. PLoS One 2019;14(1):e0209064 [FREE Full text] [doi: 10.1371/journal.pone.0209064] [Medline: 30673710]

52. Hopfer S, Hecht M, Ray A, Miller-Day M, BeLue R, Zimet G. Feasibility of implementing a community clinic based interactive health kiosk about HPV vaccination targeting African American young adult women attending Planned Parenthood. Implement Sci 2017;12(Supplement 1):S95.

53. Hopfer S, Ray AE, Hecht ML, Miller-Day M, Belue R, Zimet G, et al. Taking an HPV vaccine research-tested intervention to scale in a clinical setting. Transl Behav Med 2018;8(5):745-752 [FREE Full text] [doi: 10.1093/tbm/ibx066] [Medline: 29425333]

54. Joshi A, Amadi C, Meza J, Aguirre T, Wilhelm S. Comparison of socio-demographic characteristics of a computer based breastfeeding educational intervention among rural Hispanic women. J Community Health 2015;40(5):993-1001. [doi: 10.1007/s10900-015-0023-3] [Medline: 25868495]

55. Joshi A, Amadi C, Meza J, Aguire T, Wilhelm S. Evaluation of a computer-based bilingual breastfeeding educational program on breastfeeding knowledge, self-efficacy and intent to breastfeed among rural Hispanic women. Int J Med Inform 2016;91:10-19. [doi: 10.1016/j.ijmedinf.2016.04.001] [Medline: 27185505]

56. Gilbert JE, Howell D, King S, Sawka C, Hughes E, Angus H, et al. Quality improvement in cancer symptom assessment and control: the Provincial Palliative Care Integration Project (PPCIP). J Pain Symptom Manage 2012;43(4):663-678 [FREE Full text] [doi: 10.1016/j.jpainsymman.2011.04.028] [Medline: 22464352]

57. Pack J. Using self-service kiosks with your check-in staff. The right combination for higher patient care. Health Manag Technol 2014;35(12):14. [Medline: 25630118]

58. Ong SW, Jassal SV, Porter E, Logan AG, Miller JA. Using an electronic self-management tool to support patients with chronic kidney disease (CKD): a CKD clinic self-care model. Semin Dial 2013;26(2):195-202. [doi: 10.1111/sdi.12054] [Medline: 23406283]

59. Ong S, Min K, Porter E, Jassal V, Logan V, Miller J. Qualitative evaluation of a patient self-management kiosk use in advanced chronic kidney disease (CKD) for a 3-year period. Am J Kidney Dis 2014;63(5):A86. [doi: 10.1053/j.ajkd.2014.01.286]

60. Sun BC, Knapp H, Shamouelian A, Golden J, Goetz MB, Asch SM. Effect of an education kiosk on patient knowledge about rapid HIV screening. J Telemed Telecare 2010;16(3):158-161. [doi: 10.1258/jtt.2009.090815] [Medline: 20386037]

61. Saifu HN, Shamouelian A, Davis LG, Santana-Rios E, Goetz MB, Asch SM, et al. Impact of a kiosk educational module on HIV screening rates and patient knowledge. J Telemed Telecare 2011;17(8):446-450. [doi: 10.1258/jtt.2011.110415] [Medline: 21967998]

62. Cohen AN, Chinman MJ, Hamilton AB, Whelan F, Young AS. Using patient-facing kiosks to support quality improvement at mental health clinics. Med Care 2013;51(3 Suppl 1):S13-S20 [FREE Full text] [doi: 10.1097/MLR.0b013e31827da859] [Medline: 23407006]

63. Young AS, Cohen AN, Hamilton AB, Hellemann G, Reist C, Whelan F. Implementing patient-reported outcomes to improve the quality of care for weight of patients with schizophrenia. J Behav Health Serv Res 2019;46(1):129-139. [doi: 10.1007/s11414-018-9641-8] [Medline: 30465314]

64. Goldstein J. Private practice outcomes: validated outcomes data collection in private practice. Clin Orthop Relat Res 2010;468(10):2640-2645 [FREE Full text] [doi: 10.1007/s11999-010-1397-2] [Medline: 20532720]

65. Mosher ZA, Hudson PW, Lee SR, Perez JL, Arguello AM, McGwin Jr G, et al. Check-in kiosks in the outpatient clinical setting: fad or the future? South Med J 2020;113(3):134-139. [doi: 10.14423/SMJ.0000000000001078] [Medline: 32123929]

66. Gaydos CA, Solis M, Hsieh YH, Jett-Goheen M, Nour S, Rothman RE. Use of tablet-based kiosks in the emergency department to guide patient HIV self-testing with a point-of-care oral fluid test. Int J STD AIDS 2013;24(9):716-721 [FREE Full text] [doi: 10.1177/0956462413487321] [Medline: 23970610]

67. Hsieh YH, Gauvey-Kern M, Peterson S, Woodfield A, Deruggiero K, Gaydos CA, et al. An emergency department registration kiosk can increase HIV screening in high risk patients. J Telemed Telecare 2014;20(8):454-459 [FREE Full text] [doi: 10.1177/1357633X14555637] [Medline: 25316041]

68. Rothman RE, Gauvey-Kern M, Woodfield A, Peterson S, Tizenberg B, Kennedy J, et al. Streamlining HIV testing in the emergency department-leveraging kiosks to provide true universal screening: a usability study. Telemed J E Health 2014;20(2):122-127 [FREE Full text] [doi: 10.1089/tmj.2013.0045] [Medline: 24205808]

69. Hsieh YH, Holtgrave DR, Peterson S, Gaydos CA, Rothman RE. Novel emergency department registration kiosk for HIV screening is cost-effective. AIDS Care 2016;28(4):483-486 [FREE Full text] [doi: 10.1080/09540121.2015.1099603] [Medline: 26477440]

XSL·FO
RenderX

70.   Haukoos JS, Hopkins E, Bender B, Al-Tayyib A, Long J, Harvey J, Denver Emergency Department HIV Testing Research Consortium. Use of kiosks and patient understanding of opt-out and opt-in consent for routine rapid human immunodeficiency virus screening in the emergency department. Acad Emerg Med 2012;19(3):287-293 [FREE Full text] [doi: 10.1111/j.1553-2712.2012.01290.x] [Medline: 22435861]

71.   Smith SE, Ludwig JT, Chinchilli VM, Mehta K, Stoute JA. Use of telemedicine to diagnose tinea in Kenyan schoolchildren. Telemed J E Health 2013;19(3):166-168. [doi: 10.1089/tmj.2012.0085] [Medline: 23356383]

72.   Boltin N, Valdes D, Culley JM, Valafar H. Mobile decision support tool for emergency departments and mass casualty incidents (EDIT): initial study. JMIR Mhealth Uhealth 2018;6(6):e10727 [FREE Full text] [doi: 10.2196/10727] [Medline: 29934288]

73.   Coyle N, Kennedy A, Schull MJ, Kiss A, Hefferon D, Sinclair P, et al. The use of a self-check-in kiosk for early patient identification and queuing in the emergency department. CJEM 2019;21(6):789-792. [doi: 10.1017/cem.2019.349] [Medline: 31057137]

74.   Mahmood A, Wyant DK, Kedia S, Ahn S, Powell MP, Jiang Y, et al. Self-check-in kiosks utilization and their association with wait times in emergency departments in the United States. J Emerg Med 2020;58(5):829-840. [doi: 10.1016/j.jemermed.2019.11.019] [Medline: 31924466]

75.   Kripalani S, Hart K, Schaninger C, Bracken S, Lindsell C, Boyington DR. Use of a tablet computer application to engage patients in updating their medication list. Am J Health Syst Pharm 2019;76(5):293-300 [FREE Full text] [doi: 10.1093/ajhp/zxy047] [Medline: 30753287]

76.   Teolis MG. A MedlinePlus kiosk promoting health literacy. J Consum Health Internet 2010;14(2):126-137 [FREE Full text] [doi: 10.1080/15398281003780966] [Medline: 20808715]

77.   Lowe C, Cummin D. The use of kiosk technology in general practice. J Telemed Telecare 2010;16(4):201-203. [doi: 10.1258/jtt.2010.004011] [Medline: 20511575]

78.   Dirocco DN, Day SC. Obtaining patient feedback at point of service using electronic kiosks. Am J Manag Care 2011;17(7):e270-e276 [FREE Full text] [Medline: 21819174]

79.   McMullen KD, McConnaughy RP, Riley RA. Outreach to improve patient education at South Carolina free medical clinics. J Consum Health Internet 2011;15(2):117-131 [FREE Full text] [doi: 10.1080/15398285.2011.572779] [Medline: 22084623]

80.   Ng G, Tan N, Bahadin J, Shum E, Tan SW. Development of an automated healthcare kiosk for the management of chronic disease patients in the primary care setting. J Med Syst 2016;40(7):169. [doi: 10.1007/s10916-016-0529-y] [Medline: 27240840]

81.   Sousa P, Rodrigues J, Brandão P. HEALTH KIOSK: What factors influence the decision on how and when to use it? In: Sheerin F, Iglesias SM, Sánchez JM, Paans W, editors. E-health and standardised nursing languages: supporting practice. advancing science. Dublin, Ireland: Acendio; 2017:209-212.

82.   Bahadin J, Shum E, Ng G, Tan N, Sellayah P, Tan SW. Follow-up consultation through a healthcare kiosk for patients with stable chronic disease in a primary care setting: a prospective study. J Gen Intern Med 2017;32(5):534-539 [FREE Full text] [doi: 10.1007/s11606-016-3931-8] [Medline: 27943039]

83.   Eaton CB, Parker DR, Borkan J, McMurray J, Roberts MB, Lu B, et al. Translating cholesterol guidelines into primary care practice: a multimodal cluster randomized trial. Ann Fam Med 2011;9(6):528-537 [FREE Full text] [doi: 10.1370/afm.1297] [Medline: 22084264]

84.   Gleason-Comstock JA, Streater A, Jen KL, Artinian NT, Timmins J, Baker S, et al. Consumer health information technology in an adult public health primary care clinic: a heart health education feasibility study. Patient Educ Couns 2013;93(3):464-471. [doi: 10.1016/j.pec.2013.07.010] [Medline: 23948646]

85.   Pendleton BF, Labuda Schrop S, Ritter C, Kinion ES, McCord G, Cray JJ, et al. Underserved patients' choice of kiosk-based preventive health information. Fam Med 2010;42(7):488-495. [Medline: 20628922]

86.   Leijon M, Arvidsson D, Nilsen P, Stark Ekman D, Carlfjord S, Andersson A, et al. Improvement of physical activity by a kiosk-based electronic screening and brief intervention in routine primary health care: patient-initiated versus staff-referred. J Med Internet Res 2011;13(4):e99. [doi: 10.2196/jmir.1745] [Medline: 22107702]

87.   Chung CF, Munson SA, Thompson MJ, Baldwin LM, Kaplan J, Cline R, et al. Implementation of a new kiosk technology for blood pressure management in a family medicine clinic: from the WWAMI region practice and research network. J Am Board Fam Med 2016;29(5):620-629 [FREE Full text] [doi: 10.3122/jabfm.2016.05.160096] [Medline: 27613795]

88.   Tompson A, Fleming S, Lee MM, Monahan M, Jowett S, McCartney D, et al. Mixed-methods feasibility study of blood pressure self-screening for hypertension detection. BMJ Open 2019;9(5):e027986 [FREE Full text] [doi: 10.1136/bmjopen-2018-027986] [Medline: 31147366]

89.   Weaver NL, Nansel TR, Williams J, Tse J, Botello-Harbaum M, Willson K. Reach of a kiosk-based pediatric injury prevention program. Transl Behav Med 2011;1(4):515-522 [FREE Full text] [doi: 10.1007/s13142-011-0066-7] [Medline: 23667402]

90.   Brixey SN, Weaver NL, Guse CE, Zimmermann H, Williams J, Corden TE, et al. The impact of behavioral risk assessments and tailored health information on pediatric injury. Clin Pediatr (Phila) 2014;53(14):1383-1389. [doi: 10.1177/0009922814549544] [Medline: 25189696]

91.   Alpert BS. Are kiosk blood pressure readings trustworthy? Blood Press Monit 2012;17(6):257-258. [doi: 10.1097/MBP.0b013e32835b9ea1] [Medline: 23147536]

92.   Houle SK, Chuck AW, Tsuyuki RT. Blood pressure kiosks for medication therapy management programs: business opportunity for pharmacists. J Am Pharm Assoc (2003) 2012;52(2):188-194. [doi: 10.1331/JAPhA.2012.11217] [Medline: 22370382]

93.   Padwal RS, Townsend RR, Trudeau L, Hamilton PG, Gelfer M. Comparison of an in-pharmacy automated blood pressure kiosk to daytime ambulatory blood pressure in hypertensive subjects. J Am Soc Hypertens 2015;9(2):123-129. [doi: 10.1016/j.jash.2014.11.004] [Medline: 25600420]

94.   Langley CA, Bush J, Patel A. An evaluation: the implementation and impact of healthy living pharmacies within the Heart of Birmingham. Birmingham, UK: Aston University; 2014.

95.   Serafico ME. 114: Utilization of an automated health kiosk in estimating the incidence of overweight and obesity: an alternative technique for community-based health-risk assessment. BMJ Open 2015;5:114. [doi: 10.1136/bmjopen-2015-forum2015abstracts.114]

96.   Niemöller S, Hübner U, Egbert N, Babitsch B. How to access personal health records? Measuring the intention to use and the perceived usefulness of two different technologies: a randomised controlled study. Stud Health Technol Inform 2019;267:197-204. [doi: 10.3233/SHTI190827] [Medline: 31483273]

97.   Demiris G, Thompson H, Boquet J, Le T, Chaudhuri S, Chung J. Older adults' acceptance of a community-based telehealth wellness system. Inform Health Soc Care 2013;38(1):27-36 [FREE Full text] [doi: 10.3109/17538157.2011.647938] [Medline: 22571733]

98.   Demiris G, Thompson HJ, Reeder B, Wilamowska K, Zaslavsky O. Using informatics to capture older adults' wellness. Int J Med Inform 2013;82(11):e232-e241 [FREE Full text] [doi: 10.1016/j.ijmedinf.2011.03.004] [Medline: 21482182]

99.   Lewis MK, Hsieh YH, Gaydos CA, Peterson SC, Rothman RE. Informed consent for opt-in HIV testing via tablet kiosk: an assessment of patient comprehension and acceptability. Int J STD AIDS 2017;28(13):1292-1298 [FREE Full text] [doi: 10.1177/0956462417701009] [Medline: 28345392]

100. Orlando MS, Rothman RE, Woodfield A, Gauvey-Kern M, Peterson S, Hill PM, et al. Kiosks as tools for health information sharing: exploratory analysis of a novel ED program. Am J Emerg Med 2014;32(7):797-799 [FREE Full text] [doi: 10.1016/j.ajem.2014.04.025] [Medline: 24833098]

101. Hsieh Y, Beck KJ, Rothman RE, Gauvey-Kern M, Woodfield A, Peterson S, et al. Factors associated with patients who prefer HIV self-testing over health professional testing in an emergency department-based rapid HIV screening program. Int J STD AIDS 2017;28(11):1124-1129 [FREE Full text] [doi: 10.1177/0956462416689629] [Medline: 28114880]

102. Orlando MS, Rothman RE, Woodfield A, Gauvey-Kern M, Peterson S, Miller T, et al. Public health information delivery in the emergency department: analysis of a kiosk-based program. J Emerg Med 2016;50(2):223-227 [FREE Full text] [doi: 10.1016/j.jemermed.2015.06.075] [Medline: 26403985]

103. Lesselroth B, Adams K, Tallett S, Ragland S, Church V, Borycki EM, et al. Usability evaluation of a medication reconciliation and allergy review (MRAR) kiosk: a methodological approach for analyzing user interactions. Stud Health Technol Inform 2015;218:61-67. [Medline: 26262528]

104. Lesselroth B, Adams S, Felder R, Dorr DA, Cauthers P, Church V, et al. Using consumer-based kiosk technology to improve and standardize medication reconciliation in a specialty care setting. Jt Comm J Qual Patient Saf 2009;35(5):264-270. [doi: 10.1016/s1553-7250(09)35037-0] [Medline: 19480380]

105. Lesselroth BJ, Felder RS, Adams SM, Cauthers PD, Dorr DA, Wong GJ, et al. Design and implementation of a medication reconciliation kiosk: the automated patient history intake device (APHID). J Am Med Inform Assoc 2009;16(3):300-304 [FREE Full text] [doi: 10.1197/jamia.M2642] [Medline: 19261949]

106. Lesselroth BJ, Holahan PJ, Adams K, Sullivan ZZ, Church VL, Woods S, et al. Primary care provider perceptions and use of a novel medication reconciliation technology. Inform Prim Care 2011;19(2):105-118 [FREE Full text] [doi: 10.14236/jhi.v19i2.802] [Medline: 22417821]

107. Joshi A, Perin DM, Amadi C, Trout K. Evaluating the usability of an interactive, bi-lingual, touchscreen-enabled breastfeeding educational programme: application of Nielson's heuristics. J Innov Health Inform 2015;22(2):265-274 [FREE Full text] [doi: 10.14236/jhi.v22i2.71] [Medline: 26245240]

108. Sano M, Egelko S, Donohue M, Ferris S, Kaye J, Hayes TL, Alzheimer Disease Cooperative Study Investigators. Developing dementia prevention trials: baseline report of the home-based assessment study. Alzheimer Dis Assoc Disord 2013;27(4):356-362 [FREE Full text] [doi: 10.1097/WAD.0b013e3182769c05] [Medline: 23151596]

109. Sano M, Egelko S, Ferris S, Kaye J, Hayes TL, Mundt JC, et al. Pilot study to show the feasibility of a multicenter trial of home-based assessment of people over 75 years old. Alzheimer Dis Assoc Disord 2010;24(3):256-263 [FREE Full text] [doi: 10.1097/WAD.0b013e3181d7109f] [Medline: 20592583]

110. Shields WC, McDonald EM, McKenzie L, Wang MC, Walker AR, Gielen AC. Using the pediatric emergency department to deliver tailored safety messages: results of a randomized controlled trial. Pediatr Emerg Care 2013;29(5):628-634 [FREE Full text] [doi: 10.1097/PEC.0b013e31828e9cd2] [Medline: 23603653]

111. Shields WC, McDonald EM, McKenzie LB, Gielen AC. Does health literacy level influence the effectiveness of a kiosk-based intervention delivered in the pediatric emergency department? Clin Pediatr (Phila) 2016;55(1):48-55. [doi: 10.1177/0009922815602889] [Medline: 26333526]

112. World bank country and lending groups. World Bank. 2021. URL: https://datahelpdesk.worldbank.org/knowledgebase/articles/906519-world-bank-country-and-lending-groups [accessed 2021-02-25]

113. Improving access, responding to patients: a 'how-to' guide for GP practices. National Health Service. 2009. URL: http://data.parliament.uk/DepositedPapers/Files/DEP2009-2798/DEP2009-2798.pdf [accessed 2022-03-23]

114. Patient check-in - automated arrival. Egton. URL: https://www.egton.net/all-services/patient-check-in/ [accessed 2020-09-11]

115. Appointments in general practice, October 2018. National Health Service Digital. 2018. URL: https://digital.nhs.uk/data-and-information/publications/statistical/appointments-in-general-practice/oct-2018 [accessed 2020-09-13]

116. Wilamowska K, Le T, Demiris G, Thompson H. Using commercially available tools for multifaceted health assessment: data integration lessons learned. Comput Inform Nurs 2013;31(7):329-334 [FREE Full text] [doi: 10.1097/NXN.0b013e318295e58f] [Medline: 23728444]

117. Courtney KL, Matthews JT, McMillan JM, Person Mecca L, Smailagic A, Siewiorek D. Usability testing of a prototype multi-user telehealth kiosk. Stud Health Technol Inform 2015;208:109-113. [Medline: 25676957]

118. Koppar AR, Sridhar V. Tele-health medical diagnostics system with integrated electronic health records. Indian J Public Health Res Dev 2012;3(1):49-52.

119. Resnick HE, Ilagan PR, Kaylor MB, Mehling D, Alwan M. TEAhM-technologies for enhancing access to health management: a pilot study of community-based telehealth. Telemed J E Health 2012;18(3):166-174. [doi: 10.1089/tmj.2011.0122] [Medline: 22364270]

120. Ahn HS, Kuo IH, Datta C, Stafford R, Kerse N, Peri K, et al. Design of a kiosk type healthcare robot system for older people in private and public places. In: International Conference on Simulation, Modeling, and Programming for Autonomous Robots. 2014 Presented at: SIMPAR '14; October 20-23, 2014; Bergamo, Italy p. 578-589. [doi: 10.1007/978-3-319-11900-7_49]

121. Thompson HJ, Demiris G, Rue T, Shatil E, Wilamowska K, Zaslavsky O, et al. A Holistic approach to assess older adults' wellness using e-health technologies. Telemed J E Health 2011;17(10):794-800 [FREE Full text] [doi: 10.1089/tmj.2011.0059] [Medline: 22011052]

122. Brys S. Kiosk-based "office" extends reach of health services providers. Behav Healthc 2013;33(6):42-44. [Medline: 24494345]

123. Mudumba R. HealthSpot: analysis of a bankruptcy in kiosk-based telehealth. Telehealth Med Today 2018;2(3):11. [doi: 10.30953/tmt.v2.11]

124. Ackerman SL, Tebb K, Stein JC, Frazee BW, Hendey GW, Schmidt LA, et al. Benefit or burden? A sociotechnical analysis of diagnostic computer kiosks in four California hospital emergency departments. Soc Sci Med 2012;75(12):2378-2385. [doi: 10.1016/j.socscimed.2012.09.013] [Medline: 23063214]

125. Singer J. Bring drug dispensing into the modern age with vending machines. American Council on Science and Health. 2020. URL: https://www.acsh.org/news/2020/03/10/bring-drug-dispensing-modern-age-vending-machines-14627 [accessed 2021-03-02]

126. 5 kiosks to put in your pharmacy right now. PBA Health. URL: https://www.pbahealth.com/5-kiosks-put-pharmacy-right-now/ [accessed 2021-03-02]

127. Letafatnejad M, Maleki M, Ebrahimi P. Barriers and facilitators of deploying health kiosk in Iran: a qualitative study. J Educ Health Promot 2020;9:95 [FREE Full text] [doi: 10.4103/jehp.jehp_548_19] [Medline: 32509903]

128. World Bank and WHO: half the world lacks access to essential health services, 100 million still pushed into extreme poverty because of health expenses. World Health Organization. 2017. URL: https://www.who.int/news/item/13-12-2017-world-bank-and-who-half-the-world-lacks-access-to-essential-health-services-100-million-still-pushed-into-extreme-poverty-because-of-health-expenses [accessed 2021-03-04]

129. Global self-service tech markets on the rise; galloping CAGR in kiosk market. BCC Research. 2016. URL: https://www.bccresearch.com/pressroom/ift/global-self-service-tech-markets-on-the-rise-galloping-cagr-in-kiosk-market [accessed 2021-02-27]

130. NHSX. Digital technology assessment criteria (DTAC). National Health Service. 2020. URL: https://www.nhsx.nhs.uk/key-tools-and-info/digital-technology-assessment-criteria-dtac/ [accessed 2021-03-03]

131. Letafat-Nejad M, Ebrahimi P, Maleki M, Aryankhesal A. Utilization of integrated health kiosks: a systematic review. Med J Islam Repub Iran 2020;34:114 [FREE Full text] [doi: 10.34171/mjiri.34.114] [Medline: 33315998]

132. Rosenbek Minet L, Hansen LW, Pedersen CD, Titlestad IL, Christensen JK, Kidholm K, et al. Early telemedicine training and counselling after hospitalization in patients with severe chronic obstructive pulmonary disease: a feasibility study. BMC Med Inform Decis Mak 2015;15:3 [FREE Full text] [doi: 10.1186/s12911-014-0124-4] [Medline: 25886014]

133. Statton S, Jones R, Thomas M, North T, Endacott R, Frost A, et al. Professional learning needs in using video calls identified through workshops. BMC Med Educ 2016;16:140 [FREE Full text] [doi: 10.1186/s12909-016-0657-6] [Medline: 27165431]

134. Williamson C. Electronic self check-in for patients: the costs and consequences. Br J Gen Pract 2016;66(644):145 [FREE Full text] [doi: 10.3399/bjgp16X684025] [Medline: 26917642]

XSL•FO
RenderX

135. Lavigueur N. Concern over data protection of new Huddersfield hospital check-in system. YorkshireLive. 2014. URL: https://www.examinerlive.co.uk/news/west-yorkshire-news/concern-over-data-protection-new-7106768 [accessed 2020-09-13]

136. Coping with COVID-19: AI provider launches touch-free interface for kiosks, PC applications. Kiosk Marketplace. 2020. URL: https://www.kioskmarketplace.com/news/ai-provider-launches-touch-free-interface-for-kiosks-and-pc-applications/ [accessed 2021-03-03]

137. Maras E. Coping with COVID-19: following COVID-19: how coronavirus pandemic is impacting kiosks. Kiosk Marketplace. 2022. URL: https://www.kioskmarketplace.com/articles/following-covid-19-how-coronavirus-pandemic-is-impacting-kiosks/ [accessed 2021-03-03]

138. Borland S. mHealth Insight. 2016. URL: https://mhealthinsight.com/2016/04/14/nhs-hospital-ae-kiosks-look-like-a-waste-of-money-time/ [accessed 2021-03-04]

139. Takyi H, Watzlaf V, Matthews JT, Zhou L, Dealmeida D. Privacy and security in multi-user health kiosks. Int J Telerehabil 2017;9(1):3-14 [FREE Full text] [doi: 10.5195/ijt.2017.6217] [Medline: 28814990]

Abbreviations

BP: blood pressure
ED: emergency department
GP: general practitioner
NHS: National Health Service
PRISMA: Preferred Reporting Items for Systematic Reviews and Meta-Analyses

Viewpoint

A Roadmap for Boosting Model Generalizability for Predicting Hospital Encounters for Asthma

Gang Luo[1], DPhil

Department of Biomedical Informatics and Medical Education, University of Washington, Seattle, WA, United States

Corresponding Author:
Gang Luo, DPhil
Department of Biomedical Informatics and Medical Education
University of Washington
UW Medicine South Lake Union
850 Republican Street, Building C, Box 358047
Seattle, WA, 98195
United States
Phone: 1 206 221 4596
Fax: 1 206 221 2671
Email: gangluo@cs.wisc.edu

Abstract

In the United States, ~9% of people have asthma. Each year, asthma incurs high health care cost and many hospital encounters covering 1.8 million emergency room visits and 439,000 hospitalizations. A small percentage of patients with asthma use most health care resources. To improve outcomes and cut resource use, many health care systems use predictive models to prospectively find high-risk patients and enroll them in care management for preventive care. For maximal benefit from costly care management with limited service capacity, only patients at the highest risk should be enrolled. However, prior models built by others miss >50% of true highest-risk patients and mislabel many low-risk patients as high risk, leading to suboptimal care and wasted resources. To address this issue, 3 site-specific models were recently built to predict hospital encounters for asthma, gaining up to >11% better performance. However, these models do not generalize well across sites and patient subgroups, creating 2 gaps before translating these models into clinical use. This paper points out these 2 gaps and outlines 2 corresponding solutions: (1) a new machine learning technique to create cross-site generalizable predictive models to accurately find high-risk patients and (2) a new machine learning technique to automatically raise model performance for poorly performing subgroups while maintaining model performance on other subgroups. This gives a roadmap for future research.

Introduction

Asthma Care Management and Our Prior Work on Predictive Modeling

In the United States, ~9% of people have asthma [1-3]. Each year, asthma incurs US$ 56 billion of health care cost [4] and many hospital encounters covering 1.8 million emergency room visits and 439,000 hospitalizations [1]. As is the case with many chronic diseases, a small percentage of patients with asthma use most health care resources [5,6]. The top 1% of patients spend 25% of the health care costs. The top 20% spend 80% [5,7]. An effective approach is urgently in need to prospectively identify high-risk patients and intervene early to avoid health decline, improve outcomes, and cut resource use. Most major employers purchase and nearly all private health plans offer care management services for preventive care [8-10]. Care management is a collaborative process to assess, coordinate, plan, implement, evaluate, and monitor the services and options to meet the health and service needs of a patient [11]. A care management program employs care managers to call patients regularly to assess their status, arrange doctor appointments, and coordinate health-related services. Proper use of care management can cut down hospital encounters by up to 40% [10,12-17]; lower health care cost by up to 15% [13-18]; and improve patient satisfaction, quality of life, and adherence to treatment by 30%-60% [12]. Care management can cost >US$

5000 per patient per year [13] and normally enrolls no more than 3% of patients [7] owing to resource limits.

Correctly finding high-risk patients to enroll is crucial for effective care management. Currently, the best method to identify high-risk patients is to use models to predict each patient's risk [19]. Many health plans such as those in 9 of 12 metropolitan communities [20] and many health care systems [21] use this method for care management. For patients predicted to have the highest risk, care managers manually review patients' medical records, consider factors such as social dimensions, and make enrollment decisions. However, prior models built by others miss >50% of true highest-risk patients and mislabel many low-risk patients as high risk [5,12,22-36]. This makes enrollment align poorly with patients who would benefit most from care management [12], leading to suboptimal care and higher costs. As the patient population is large, a small boost in model performance will benefit many patients and produce a large positive impact. Of the top 1% patients with asthma who would incur the highest costs, for every 1% more whom one could find and enroll, one could save up to US$ 21 million more in asthma care every year as well as improve outcomes [5,26,27].

To address the issue of low model performance, we recently built 3 site-specific models to predict whether a patient with asthma would incur any hospital encounter for asthma in the subsequent 12 months, 1 model for each of the 3 health care systems—the University of Washington Medicine (UWM), Intermountain Healthcare (IH), and Kaiser Permanente Southern California (KPSC) [21,37,38]. Each prior model that others built for a comparable outcome [5,26-34] had an area under the receiver operating characteristic curve (AUC) that was ≤0.79 and a sensitivity that was ≤49%. Our models raised the AUC to 0.9 and the sensitivity to 70% on UWM data [21], the AUC to 0.86 and the sensitivity to 54% on IH data [37], and the AUC to 0.82 and the sensitivity to 52% on KPSC data [38].

Our eventual goal is to translate our models into clinical use. However, despite major progress, our models do not generalize well across sites and patient subgroups, and 2 gaps remain.

Gap 1: The Site-Specific Models Have Suboptimal Generalizability When Applied to the Other Sites

Each of our models was built for 1 site. As is typical in predictive modelling [39,40], when applied to the other sites, the site-specific model had AUC drops of up to 4.1% [38], potentially degrading care management enrollment decisions. One can do transfer learning using other source health care systems' raw data to boost model performance for the target health care system [41-45], but health care systems are seldom willing to share raw data. Research networks [46-48] mitigate the problem but do not solve it. Many health care systems are not in any network. Health care systems in the network share raw data of finite attributes. Our prior model-based transfer learning approach [49] requires no raw data from other health care systems. However, it does not control the number of features (independent variables) used in the final model for the target site, creating difficulty to build the final model for the target site for clinical use. Consequently, it is never implemented in computer code.

Gap 2: The Models Exhibit Large Performance Gaps When Applied to Specific Patient Subgroups

Our models performed up to 8% worse on Black patients. This is a typical barrier in machine learning, where many models exhibit large subgroup performance gaps, for example, of up to 38% [50-57]. No existing tool for auditing model bias and fairness [58,59] has been applied to our models. Currently, it is unknown how our models perform on key patient subgroups defined by independent variables such as race, ethnicity, and insurance type. In other words, it is unknown how our models perform for different races, different ethnicities, and patients using different types of insurance. Large performance gaps among patient subgroups can lead to care inequity and should be avoided.

Many methods to improve fairness in machine learning exist [50-52]. These methods usually boost model performance on some subgroups at the price of lowering both model performance on others and the overall model performance [50-52]. Lowering the overall model performance is undesired [51,57]. Owing to the large patient population, even a 1% drop in the overall model performance could potentially degrade many patients' outcomes. Chen et al [57] cut model performance gaps among subgroups by collecting more training data and adding additional features, both of which are often difficult or infeasible to do. For classifying images via machine learning, Goel et al's method [55] raised the overall model performance and cut model performance gaps among subgroups of a value of the dependent variable—not among subgroups defined by independent variables. The dependent variable is also known as the outcome or the prediction target. An example of the dependent variable is whether a patient with asthma will incur any hospital encounter for asthma in the subsequent 12 months. The independent variables are also known as features. Race, ethnicity, and insurance type are 3 examples of independent variables. Many machine learning techniques to handle imbalanced classes exist [60,61]. In these techniques, subgroups are defined by the dependent variable rather than by independent variables.

Contributions of This Paper

To fill the 2 gaps on suboptimal model generalizability and let more high-risk patients obtain appropriate and equitable preventive care, the paper makes 2 contributions, thereby giving a roadmap for future research.

1. To address the first gap, a new machine learning technique is outlined to create cross-site generalizable predictive models to accurately find high-risk patients. This is to cut model performance drop across sites.

2. To address the second gap, a new machine learning technique is outlined to automatically raise model performance for poorly performing subgroups while maintaining model performance on other subgroups. This is to cut model performance gaps among patient subgroups and to reduce care inequity.

The following sections describe the main ideas of the proposed new machine learning techniques.

Machine Learning Technique for Creating Cross-Site Generalizable Predictive Models to Accurately Find High-risk Patients

Our Prior Models

In our prior work [21,37,38], for each of the 3 health care systems (sites), namely, KPSC, IH, and UWM, >200 candidate features were checked and the site's data were used to build a full site-specific extreme gradient boosting (XGBoost) model to predict hospital encounters for asthma. XGBoost [62] automatically chose the features to be used in the model from the candidate features, computed their importance values, and ranked them in the descending order of these values. The top (~20) features with importance values ≥1% have nearly all of the predictive power of all (on average ~140) features used in the model [21,37,38]. Although some lower-ranked features are unavailable at other sites, each top feature such as the number of patient's asthma-related emergency room visits in the prior 12 months is computed using (eg, diagnosis, encounter) attributes routinely collected by almost every American health care system that uses electronic medical records. Using the top features and the site's data, a simplified XGBoost model was built. It, but not the full model, can be applied to other sites. The simplified model performed similarly to the full model at the site. However, when applied to another site, even after being retrained on its data, the simplified model performed up to 4.1% worse than the full model built specifically for it, as distinct sites have only partially overlapping top features [21,37,38].

Building Cross-Site Generalizable Models

To ensure that the same variable is called the same name at different sites and the variable's content is recorded in the same way across these sites, the data sets at all source sites and the target site are converted into the Observational Medical Outcomes Partnership (OMOP) common data model [63] and its linked standardized terminologies [64]. If needed, the data model is extended to cover the variables that are not included in the original data model but exist in the data sets.

Our goal is to build cross-site generalizable models fulfilling 2 conditions. First, the model uses a moderate number of features. Controlling the number of features used in the model would ease the future clinical deployment of the model. Second, a separate component or copy of the model is initially built at each source site. When applied to the target site and possibly after being retrained on its data, the model performs similarly to the full model built specifically for it. To reach our goal for the case of IH and UWM being the source sites and KPSC being the target site, we proceed in 2 steps (Figure 1). In step 1, the top features found at each source site are combined. For each source site, the combined top features, its data, and the machine learning algorithm adopted to build its full model are used to build an expanded simplified model. Compared with the original simplified model built for the site, the expanded simplified model uses more features with predictive power and tends to generalize better across sites. In step 2, model-based transfer learning is conducted to further boost model performance. For each data instance of the target site, each source site's expanded simplified model is applied to the data instance, a prediction result is computed, and the prediction result is used as a new feature. For the target site, its data, the combined top features found at the source sites, and the new features are used to build its final model.

To reach our goal for the case that IH or UWM is the target site and KPSC is one of the source sites, we need to address the issue that the claim-based features used at KPSC [38] are unavailable at IH, UWM, and many other health care systems with no claim data. At KPSC, these features are dropped and the other candidate features are used to build a site-specific model and recompute the top features. This helps reach the effect that the top features found at each of KPSC, IH, and UWM are available at all 3 sites and almost every other American health care system that uses electronic medical record systems. In the unlikely case that any recomputed top feature at KPSC violates this, the feature is skipped when building cross-site generalizable models.

Our method to build cross-site generalizable models can handle all kinds of prediction targets, features, and models used at the source and target sites. Given a distinct prediction target, if some top features found at a source site are unavailable at many American health care systems using electronic medical record systems, the drop→recompute→skip approach shown above can be used to handle these features. Moreover, at any source site, if the machine learning algorithm used to build the full site-specific model is like XGBoost [62] or random forest that automatically computes feature importance values, the top features with the highest importance values can be used. Otherwise, if the algorithm used to build the full model does not automatically compute feature importance values, an automatic feature selection method [65] like the information gain method can be used to choose the top features. Alternatively, XGBoost or random forest can be used to build a model, automatically compute feature importance values, and choose the top features with the highest importance values.

Our new model-based transfer learning approach waives the need for source sites' raw data. Health care systems are more willing to share with others trained models than raw data. A model trained using the data of a source site contains much information that is useful for the prediction task at the target site. This information offers much value when the target site has insufficient data for model training. If the target site is large, this information can still be valuable. Distinct sites have differing data pattern distributions. A pattern that matches a small percentage of patients and is difficult to identify at the target site could match a larger percentage of patients and be easier to identify at one of the source sites. In this case, its expanded simplified model could incorporate the pattern through model training to better predict the outcomes of certain types of patients, which is difficult to do using only the information from the target site but no information from the source sites. Thus, we expect that compared with just retraining a source site's expanded simplified model on the target site's data, doing model-based transfer learning in step 2 could lead to a better performing final model for the target site.

When the target site goes beyond IH, UWM, and KPSC, IH, UWM, and KPSC can be used as the source sites to have more

top features to combine. This would make our cross-site models generalize even better.

Figure 1. The method used in this study to build cross-site generalizable models. IH: Intermountain Healthcare. KPSC: Kaiser Permanente Southern California. UWM: University of Washington Medicine.

$$g(f_1(\vec{x}), f_2(\vec{x}), \vec{x})$$
$$g: \text{the final model for KPSC}$$

$$f_1(\vec{x}) \qquad f_2(\vec{x}) \qquad \vec{x}$$

f_1: a model trained using IH's data f_2: a model trained using UWM's data

$\vec{x}$: the top features found at IH and UWM in a KPSC data instance
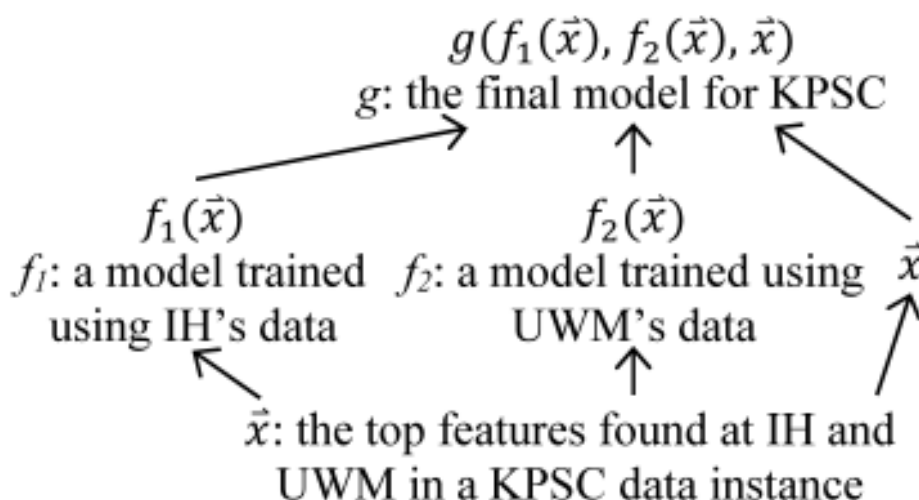
Machine Learning Technique for Automatically Raising Model Performance for Poorly Performing Patient Subgroups While Maintaining Model Performance on Other Subgroups to Reduce Care Inequity

Several clinical experts are asked to identify several patient subgroups of great interest to clinicians (eg, by race, ethnicity, insurance type) through discussion. These subgroups are not necessarily mutually exclusive of each other. Each subgroup is defined by one or more attribute values. Given a predictive model built on a training set, model performance on each subgroup on the test set is computed and shown [58,59]. Machine learning needs enough training data to work well. Often, the model performs much worse on a small subgroup than on a large subgroup [50,52]. After identifying 1 or more target subgroups where the model performs much worse than on other subgroups [51], a new dual-model approach is used to raise model performance on the target subgroups while maintaining model performance on other subgroups.

More specifically, given n target patient subgroups, they are sorted as G_i ($1 \le i \le n$) in ascending order of size and oversampled based on n integers r_i ($1 \le i \le n$) satisfying $r_1 \ge r_2 \ge \dots \ge r_n > 1$. As Figure 2 shows, for each training instance in G_1, r_1 copies of it including itself are made. For each training instance in [×] ($2 \le j \le n$), r_j copies of it, including itself, are made. Intuitively, the smaller the i ($1 \le i \le n$) and thus G_i, the more aggressive oversampling is needed on G_i for machine learning to work well on it. The sorting ensures that if a training instance appears in ≥ 2 target subgroups, copies are made for it based on the largest r_i of these subgroups. If needed, 1 set of r_i's could be used for training instances with bad outcomes, and another set of r_i's could be used for training instances with good outcomes [66].

[×] is the union of the n target subgroups. Using the training instances outside G, the copies made for the training instances in G and an automatic machine learning model selection method like our formerly developed one [67], the AUC on G is optimized, the values of r_i ($1 \le i \le n$) are automatically selected, and a second model is trained. As is typical in using oversampling to improve fairness in machine learning, compared with the original model, the second model tends to perform better on G and worse on the patients outside G [51,66] because oversampling increases the percentage of training instances in G and decreases the percentage of training instances outside G. To avoid running into the case of having insufficient data for model training, no undersampling is performed on the training instances outside G. The original model is used to make predictions on the patients outside G. The second model is used to make predictions on the patients in G. In this way, model performance on G can be raised without lowering either model performance on the patients outside G or the overall model performance. All patients' data instead of only the training instances in G are used to train the second model. Otherwise, the second model may perform poorly on G owing to insufficient training data in G [51]. For a similar reason, we choose to not use decoupled classifiers, where a separate classifier is trained for each subgroup by using only that subgroup's data [51] on the target subgroups [57].

The above discussion focuses on the case that the original model is built on 1 site's data without using any other site's information. When the original model is a cross-site generalizable model built for the target site using the method in the "Building cross-site generalizable models" section and models trained at the source sites, to raise model performance on the target patient subgroups, we change the way to build the second model for the target site by proceeding in 2 steps (Figure 3). In step 1, the top features found at each source site are combined. Recall that G is the union of the n target subgroups. For each source site, the target subgroups are oversampled in the way mentioned above; the AUC on G at the source site is

optimized; and its data both in and outside G, the combined top features, and the machine learning algorithm adopted to build its full model are used to build a second expanded simplified model. In step 2, model-based transfer learning is conducted to incorporate useful information from the source sites. For each data instance of the target site, each source site's second expanded simplified model is applied to the data instance, a prediction result is computed, and the prediction result is used

as a new feature. For the target site, the target subgroups are oversampled in the way mentioned above, the AUC on G at the target site is optimized, and its data both in and outside G, the combined top features found at the source sites, and the new features are used to build the second model for it. For each i ($1 \leq i \leq n$), each of the source and target sites could use a distinct oversampling ratio r_i.

Figure 2. Oversampling for 3 target patient subgroups G_1, G_2, and G_3.
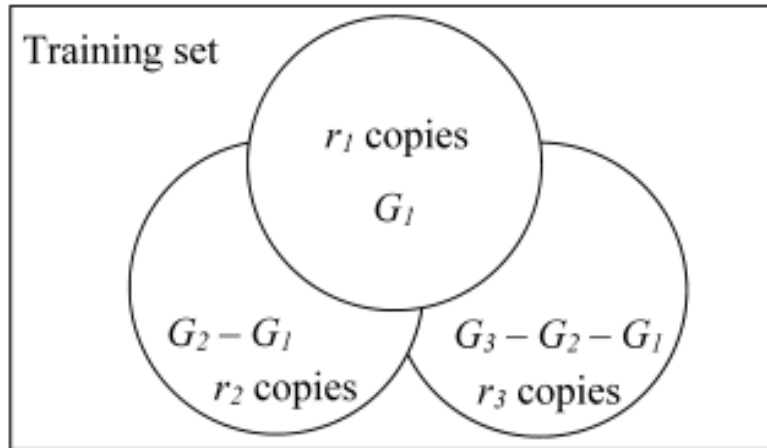
Figure 3. The method used in this study to boost a cross-site generalizable model's performance on the target patient subgroups. IH: Intermountain Healthcare. KPSC: Kaiser Permanente Southern California. UWM: University of Washington Medicine.
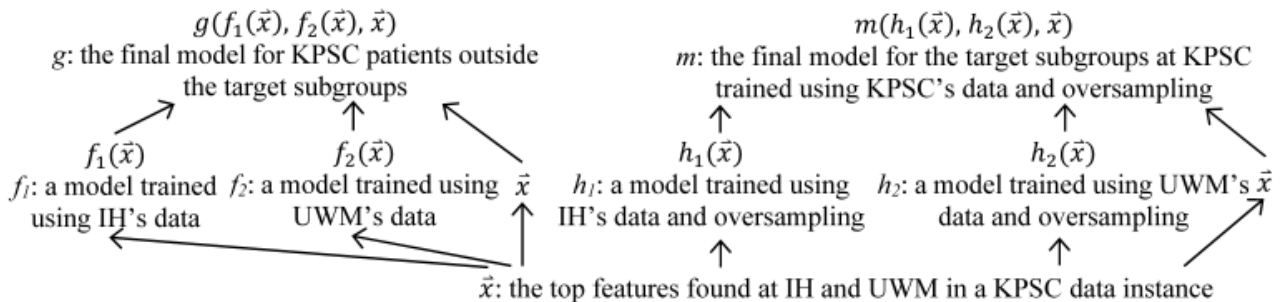
Discussion

Predictive models differ by diseases and other factors. However, our proposed machine learning techniques are general and depend on no specific disease, patient cohort, or health care system. Given a new data set with a differing prediction target, disease, patient cohort, set of health care systems, or set of variables, one can use our proposed machine learning techniques to improve model generalizability across sites, as well as to boost model performance on poorly performing patient subgroups while maintaining model performance on others. For instance, our proposed machine learning techniques can be used to improve model performance for predicting other outcomes such as adherence to treatment [68] and no-shows [69]. This will help target resources such as interventions to improve adherence to treatment [68] and reminders by phone calls to reduce no-shows [69]. Care management is widely adopted to manage patients with chronic obstructive pulmonary disease, patients with diabetes, and patients with heart disease [6], where

our proposed machine learning techniques can also be used. Our proposed predictive models are based on the OMOP common data model [63] and its linked standardized terminologies [64], which standardize administrative and clinical variables from at least 10 large health care systems in the United States [47,70]. Our proposed predictive models apply to those health care systems and others using OMOP.

Conclusions

To better identify patients likely to benefit most from asthma care management, we recently built the most accurate models to date to predict hospital encounters for asthma. However, these models do not generalize well across sites and patient subgroups, creating 2 gaps before translating these models into clinical use. This paper points out these 2 gaps and outlines 2 corresponding solutions, giving a roadmap for future research. The principles of our proposed machine learning techniques generalize to many other clinical predictive modeling tasks.

Acknowledgments

Conflicts of Interest

None declared.

References

1. FastStats asthma. Centers for Disease Control and Prevention. 2021. URL: http://www.cdc.gov/nchs/fastats/asthma.htm [accessed 2022-02-17]
2. Akinbami LJ, Moorman JE, Liu X. Asthma prevalence, health care use, and mortality: United States, 2005-2009. Natl Health Stat Report 2011 Jan 12(32):1-14 [FREE Full text] [Medline: 21355352]
3. Akinbami LJ, Moorman JE, Bailey C, Zahran HS, King M, Johnson CA, et al. Trends in asthma prevalence, health care use, and mortality in the United States, 2001-2010. NCHS Data Brief 2012 May(94):1-8 [FREE Full text] [Medline: 22617340]
4. Asthma in the US. Centers for Disease Control and Prevention. 2021. URL: http://www.cdc.gov/vitalsigns/asthma [accessed 2022-02-17]
5. Schatz M, Nakahiro R, Jones CH, Roth RM, Joshua A, Petitti D. Asthma population management: development and validation of a practical 3-level risk stratification scheme. Am J Manag Care 2004 Jan;10(1):25-32 [FREE Full text] [Medline: 14738184]
6. Duncan I. Healthcare Risk Adjustment and Predictive Modeling, Second Edition. Winsted, CT: ACTEX Publications Inc; 2018.
7. Axelrod RC, Vogel D. Predictive modeling in health plans. Disease Manage Health Outcomes 2003;11(12):779-787. [doi: 10.2165/00115677-200311120-00003]
8. Vogeli C, Shields AE, Lee TA, Gibson TB, Marder WD, Weiss KB, et al. Multiple chronic conditions: prevalence, health consequences, and implications for quality, care management, and costs. J Gen Intern Med 2007 Dec;22 Suppl 3:391-395 [FREE Full text] [doi: 10.1007/s11606-007-0322-1] [Medline: 18026807]
9. Nelson L. Lessons from Medicare's demonstration projects on disease management and care coordination. Congressional Budget Office. 2012. URL: https://www.cbo.gov/sites/default/files/112th-congress-2011-2012/workingpaper/WP2012-01_Nelson_Medicare_DMCC_Demonstrations_1.pdf [accessed 2022-02-15]
10. Caloyeras JP, Liu H, Exum E, Broderick M, Mattke S. Managing manifest diseases, but not health risks, saved PepsiCo money over seven years. Health Aff (Millwood) 2014 Jan;33(1):124-131. [doi: 10.1377/hlthaff.2013.0625] [Medline: 24395944]
11. Definition and philosophy of case management. Commission for Case Manager Certification. 2021. URL: https://ccmcertification.org/about-ccmc/about-case-management/definition-and-philosophy-case-management [accessed 2022-02-17]
12. Levine SH, Adams J, Attaway K, Dorr DA, Leung M, Popescu P, et al. Predicting the financial risks of seriously ill patients. California Health Care Foundation. 2011. URL: http://www.chcf.org/publications/2011/12/predictive-financial-risks [accessed 2022-02-17]
13. Rubin RJ, Dietrich KA, Hawk AD. Clinical and economic impact of implementing a comprehensive diabetes management program in managed care. J Clin Endocrinol Metab 1998 Aug;83(8):2635-2642. [doi: 10.1210/jcem.83.8.5075] [Medline: 9709924]
14. Greineder DK, Loane KC, Parks P. A randomized controlled trial of a pediatric asthma outreach program. J Allergy Clin Immunol 1999 Mar;103(3 Pt 1):436-440. [doi: 10.1016/s0091-6749(99)70468-9] [Medline: 10069877]
15. Kelly CS, Morrow AL, Shults J, Nakas N, Strope GL, Adelman RD. Outcomes evaluation of a comprehensive intervention program for asthmatic children enrolled in Medicaid. Pediatrics 2000 May;105(5):1029-1035. [doi: 10.1542/peds.105.5.1029] [Medline: 10790458]
16. Axelrod RC, Zimbro KS, Chetney RR, Sabol J, Ainsworth VJ. A disease management program utilizing life coaches for children with asthma. J Clin Outcomes Manag. 2001. URL: https://www.researchgate.net/publication/284394600_A_disease_management_program_utilising_life_coaches_for_children_with_asthma [accessed 2022-02-22]
17. Dorr DA, Wilcox AB, Brunker CP, Burdon RE, Donnelly SM. The effect of technology-supported, multidisease care management on the mortality and hospitalization of seniors. J Am Geriatr Soc 2008 Dec;56(12):2195-2202. [doi: 10.1111/j.1532-5415.2008.02005.x] [Medline: 19093919]
18. Beaulieu N, Cutler DM, Ho K, Isham G, Lindquist T, Nelson A, et al. The business case for diabetes disease management for managed care organizations. Forum Health Econ Policy 2006;9(1):1-37. [doi: 10.2202/1558-9544.1072]
19. Curry N, Billings J, Darin B, Dixon J, Williams M, Wennberg D. Predictive risk project literature review. London: King's Fund. 2005. URL: http://www.kingsfund.org.uk/sites/files/kf/field/field_document/predictive-risk-literature-review-june2005.pdf [accessed 2022-02-17]

20.   Mays GP, Claxton G, White J. Managed care rebound? Recent changes in health plans' cost containment strategies. Health Aff (Millwood) 2004;Suppl Web Exclusives:W4-427-W4-436. [doi: 10.1377/hlthaff.w4.427] [Medline: 15451964]

21.   Tong Y, Messinger AI, Wilcox AB, Mooney SD, Davidson GH, Suri P, et al. Forecasting future asthma hospital encounters of patients with asthma in an academic health care system: predictive model development and secondary analysis study. J Med Internet Res 2021 Apr 16;23(4):e22796 [FREE Full text] [doi: 10.2196/22796] [Medline: 33861206]

22.   Ash A, McCall N. Risk assessment of military populations to predict health care cost and utilization. Research Triangle Institute. 2005. URL: http://www.rti.org/pubs/tricare_riskassessment_final_report_combined.pdf [accessed 2022-02-17]

23.   Diehr P, Yanez D, Ash A, Hornbrook M, Lin DY. Methods for analyzing health care utilization and costs. Annu Rev Public Health 1999;20:125-144. [doi: 10.1146/annurev.publhealth.20.1.125] [Medline: 10352853]

24.   Iezzoni L. Risk Adjustment for Measuring Health Care Outcomes, Fourth Edition. Chicago, IL: Health Administration Press; 2012.

25.   Weir S, Aweh G, Clark RE. Case selection for a Medicaid chronic care management program. Health Care Financ Rev 2008;30(1):61-74 [FREE Full text] [Medline: 19040174]

26.   Schatz M, Cook EF, Joshua A, Petitti D. Risk factors for asthma hospitalizations in a managed care organization: development of a clinical prediction rule. Am J Manag Care 2003 Aug;9(8):538-547 [FREE Full text] [Medline: 12921231]

27.   Lieu TA, Quesenberry CP, Sorel ME, Mendoza GR, Leong AB. Computer-based models to identify high-risk children with asthma. Am J Respir Crit Care Med 1998 Apr;157(4 Pt 1):1173-1180. [doi: 10.1164/ajrccm.157.4.9708124] [Medline: 9563736]

28.   Lieu TA, Capra AM, Quesenberry CP, Mendoza GR, Mazar M. Computer-based models to identify high-risk adults with asthma: is the glass half empty or half full? J Asthma 1999 Jun;36(4):359-370. [doi: 10.3109/02770909909068229] [Medline: 10386500]

29.   Forno E, Fuhlbrigge A, Soto-Quirós ME, Avila L, Raby BA, Brehm J, et al. Risk factors and predictive clinical scores for asthma exacerbations in childhood. Chest 2010 Nov;138(5):1156-1165 [FREE Full text] [doi: 10.1378/chest.09-2426] [Medline: 20472862]

30.   Loymans RJB, Debray TPA, Honkoop PJ, Termeer EH, Snoeck-Stroband JB, Schermer TRJ, et al. Exacerbations in adults with asthma: a systematic review and external validation of prediction models. J Allergy Clin Immunol Pract 2018;6(6):1942-1952.e15. [doi: 10.1016/j.jaip.2018.02.004] [Medline: 29454163]

31.   Eisner MD, Yegin A, Trzaskoma B. Severity of asthma score predicts clinical outcomes in patients with moderate to severe persistent asthma. Chest 2012 Jan;141(1):58-65. [doi: 10.1378/chest.11-0020] [Medline: 21885725]

32.   Sato R, Tomita K, Sano H, Ichihashi H, Yamagata S, Sano A, et al. The strategy for predicting future exacerbation of asthma using a combination of the Asthma Control Test and lung function test. J Asthma 2009 Sep;46(7):677-682. [doi: 10.1080/02770900902972160] [Medline: 19728204]

33.   Yurk RA, Diette GB, Skinner EA, Dominici F, Clark RD, Steinwachs DM, et al. Predicting patient-reported asthma outcomes for adults in managed care. Am J Manag Care 2004 May;10(5):321-328 [FREE Full text] [Medline: 15152702]

34.   Xiang Y, Ji H, Zhou Y, Li F, Du J, Rasmy L, et al. Asthma exacerbation prediction and risk factor analysis based on a time-sensitive, attentive neural network: retrospective cohort study. J Med Internet Res 2020 Jul 31;22(7):e16981 [FREE Full text] [doi: 10.2196/16981] [Medline: 32735224]

35.   Miller MK, Lee JH, Blanc PD, Pasta DJ, Gujrathi S, Barron H, TENOR Study Group. TENOR risk score predicts healthcare in adults with severe or difficult-to-treat asthma. Eur Respir J 2006 Dec;28(6):1145-1155 [FREE Full text] [doi: 10.1183/09031936.06.00145105] [Medline: 16870656]

36.   Loymans RJ, Honkoop PJ, Termeer EH, Snoeck-Stroband JB, Assendelft WJ, Schermer TR, et al. Identifying patients at risk for severe exacerbations of asthma: development and external validation of a multivariable prediction model. Thorax 2016 Sep;71(9):838-846. [doi: 10.1136/thoraxjnl-2015-208138] [Medline: 27044486]

37.   Luo G, He S, Stone BL, Nkoy FL, Johnson MD. Developing a model to predict hospital encounters for asthma in asthmatic patients: secondary analysis. JMIR Med Inform 2020 Jan 21;8(1):e16080 [FREE Full text] [doi: 10.2196/16080] [Medline: 31961332]

38.   Luo G, Nau CL, Crawford WW, Schatz M, Zeiger RS, Rozema E, et al. Developing a predictive model for asthma-related hospital encounters in patients with asthma in a large, integrated health care system: secondary analysis. JMIR Med Inform 2020 Nov 09;8(11):e22689 [FREE Full text] [doi: 10.2196/22689] [Medline: 33164906]

39.   Bleeker SE, Moll HA, Steyerberg EW, Donders AR, Derksen-Lubsen G, Grobbee DE, et al. External validation is necessary in prediction research: a clinical example. J Clin Epidemiol 2003 Sep;56(9):826-832. [doi: 10.1016/s0895-4356(03)00207-5] [Medline: 14505766]

40.   Siontis GC, Tzoulaki I, Castaldi PJ, Ioannidis JP. External validation of new risk prediction models is infrequent and reveals worse prognostic discrimination. J Clin Epidemiol 2015 Jan;68(1):25-34. [doi: 10.1016/j.jclinepi.2014.09.007] [Medline: 25441703]

41.   Wiens J, Guttag J, Horvitz E. A study in transfer learning: leveraging data from multiple hospitals to enhance hospital-specific predictions. J Am Med Inform Assoc 2014;21(4):699-706 [FREE Full text] [doi: 10.1136/amiajnl-2013-002162] [Medline: 24481703]

XSL•FO
RenderX

42.   Gong JJ, Sundt TM, Rawn JD, Guttag JV. Instance weighting for patient-specific risk stratification models. 2015 Presented at: ACM SIGKDD International Conference on Knowledge Discovery and Data Mining; August 10-13; Sydney, NSW, Australia p. 369-378. [doi: 10.1145/2783258.2783397]

43.   Lee G, Rubinfeld I, Syed Z. Adapting surgical models to individual hospitals using transfer learning. 2012 Presented at: IEEE International Conference on Data Mining Workshops; December 10; Brussels, Belgium p. 57-63. [doi: 10.1109/icdmw.2012.93]

44.   Pan S, Yang Q. A survey on transfer learning. IEEE Trans Knowl Data Eng 2010 Oct;22(10):1345-1359. [doi: 10.1109/TKDE.2009.191]

45.   Weiss K, Khoshgoftaar TM, Wang D. A survey of transfer learning. J Big Data 2016 May 28;3:9. [doi: 10.1186/s40537-016-0043-6]

46.   Jayanthi A. Down the rabbit hole at Epic: 9 key points from the users group meeting. Becker's Health IT. 2016. URL: http://www.beckershospitalreview.com/healthcare-information-technology/down-the-rabbit-hole-at-epic-8-key-points-from-the-users-group-meeting.html [accessed 2022-02-17]

47.   Hripcsak G, Duke JD, Shah NH, Reich CG, Huser V, Schuemie MJ, et al. Observational Health Data Sciences and Informatics (OHDSI): opportunities for observational researchers. Stud Health Technol Inform 2015;216:574-578 [FREE Full text] [Medline: 26262116]

48.   Fleurence RL, Curtis LH, Califf RM, Platt R, Selby JV, Brown JS. Launching PCORnet, a national patient-centered clinical research network. J Am Med Inform Assoc 2014;21(4):578-582 [FREE Full text] [doi: 10.1136/amiajnl-2014-002747] [Medline: 24821743]

49.   Luo G, Sward K. A roadmap for optimizing asthma care management via computational approaches. JMIR Med Inform 2017 Sep 26;5(3):e32 [FREE Full text] [doi: 10.2196/medinform.8076] [Medline: 28951380]

50.   Oakden-Rayner L, Dunnmon J, Carneiro G, Ré C. Hidden stratification causes clinically meaningful failures in machine learning for medical imaging. 2020 Presented at: ACM Conference on Health, Inference, and Learning; April 2-4; Toronto, Ontario, Canada p. 151-159. [doi: 10.1145/3368555.3384468]

51.   Caton S, Haas C. Fairness in machine learning: a survey. Arxiv. 2020. URL: https://arxiv.org/abs/2010.04053 [accessed 2022-02-18]

52.   Barocas S, Hardt M, Narayanan A. Fairness and Machine Learning: Limitations and Opportunities. 2021. URL: https://fairmlbook.org [accessed 2022-02-17]

53.   DeVries T, Misra I, Wang C, van der Maaten L. Does object recognition work for everyone? 2019 Presented at: IEEE Conference on Computer Vision and Pattern Recognition Workshops; June 16-20; Long Beach, CA p. 52-59. [doi: 10.1109/cvprw47913.2019]

54.   Buolamwini J, Gebru T. Gender shades: intersectional accuracy disparities in commercial gender classification. 2018 Presented at: Conference on Fairness, Accountability and Transparency; February 23-24; New York, NY p. 77-91 URL: https://proceedings.mlr.press/v81/buolamwini18a/buolamwini18a.pdf

55.   Goel K, Gu A, Li Y, Ré C. Model patching: closing the subgroup performance gap with data augmentation. 2021 Presented at: Proceedings of the 9th International Conference on Learning Representations; May 3-7; Vienna, Austria p. 1-30 URL: https://openreview.net/forum?id=9YlaeLfuhJF

56.   Seyyed-Kalantari L, Liu G, McDermott M, Chen IY, Ghassemi M. CheXclusion: Fairness gaps in deep chest X-ray classifiers. Pac Symp Biocomput 2021;26:232-243 [FREE Full text] [Medline: 33691020]

57.   Chen IY, Johansson FD, Sontag DA. Why is my classifier discriminatory? 2018 Presented at: Proceedings of Annual Conference on Neural Information Processing Systems; December 3-8; Montréal, Canada p. 3543-3554 URL: https://dl.acm.org/doi/10.5555/3327144.3327272

58.   Saleiro P, Kuester B, Stevens A, Anisfeld A, Hinkson L, London J, et al. Aequitas: a bias and fairness audit toolkit. Arxiv. 2018. URL: https://arxiv.org/abs/1811.05577 [accessed 2022-02-18]

59.   Panigutti C, Perotti A, Panisson A, Bajardi P, Pedreschi D. FairLens: Auditing black-box clinical decision support systems. Inf Process Manag 2021 Sep;58(5):102657. [doi: 10.1016/j.ipm.2021.102657]

60.   Branco P, Torgo L, Ribeiro RP. A survey of predictive modeling on imbalanced domains. ACM Comput Surv 2016 Nov 11;49(2):31. [doi: 10.1145/2907070]

61.   Kaur H, Pannu HS, Malhi AK. A systematic review on imbalanced data challenges in machine learning: applications and solutions. ACM Comput Surv 2020 Jul 31;52(4):79. [doi: 10.1145/3343440]

62.   Chen T, Guestrin C. XGBoost: A scalable tree boosting system. 2016 Presented at: Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining; August 13-17; San Francisco, CA p. 785-794. [doi: 10.1145/2939672.2939785]

63.   Data standardization. Observational Health Data Sciences and Informatics. 2021. URL: https://www.ohdsi.org/data-standardization [accessed 2022-02-17]

64.   Standardized vocabularies. Observational Health Data Sciences and Informatics. 2021. URL: https://www.ohdsi.org/web/wiki/doku.php?id=documentation:vocabulary:sidebar [accessed 2022-02-17]

65.   Witten IH, Frank E, Hall MA, Pal CJ. Data Mining: Practical Machine Learning Tools and Techniques, 4th edition. Burlington, MA: Morgan Kaufmann; 2016.

66. Rancic S, Radovanovic S, Delibasic B. Investigating oversampling techniques for fair machine learning models. 2021 Presented at: Proceedings of the 7th International Conference on Decision Support System Technology; May 26-28; Loughborough, UK p. 110-123. [doi: 10.1007/978-3-030-73976-8_9]

67. Zeng X, Luo G. Progressive sampling-based Bayesian optimization for efficient and automatic machine learning model selection. Health Inf Sci Syst 2017 Dec;5(1):2 [FREE Full text] [doi: 10.1007/s13755-017-0023-z] [Medline: 29038732]

68. Kumamaru H, Lee MP, Choudhry NK, Dong YH, Krumme AA, Khan N, et al. Using previous medication adherence to predict future adherence. J Manag Care Spec Pharm 2018 Nov;24(11):1146-1155. [doi: 10.18553/jmcp.2018.24.11.1146] [Medline: 30362915]

69. Chariatte V, Berchtold A, Akré C, Michaud PA, Suris JC. Missed appointments in an outpatient clinic for adolescents, an approach to predict the risk of missing. J Adolesc Health 2008 Jul;43(1):38-45. [doi: 10.1016/j.jadohealth.2007.12.017] [Medline: 18565436]

70. Overhage JM, Ryan PB, Reich CG, Hartzema AG, Stang PE. Validation of a common data model for active safety surveillance research. J Am Med Inform Assoc 2012;19(1):54-60 [FREE Full text] [doi: 10.1136/amiajnl-2011-000376] [Medline: 22037893]

Abbreviations

AUC: area under the receiver operating characteristic curve
IH: Intermountain Healthcare
KPSC: Kaiser Permanente Southern California
OMOP: Observational Medical Outcomes Partnership
UWM: University of Washington Medicine
XGBoost: extreme gradient boosting

XSL•FO
RenderX

Viewpoint

Primary Care: The Actual Intelligence Required for Artificial Intelligence to Advance Health Care and Improve Health

Winston R Liaw[1*], MD, MPH; John M Westfall[2*], MD, MPH; Tyler S Williamson[3*], PhD; Yalda Jabbarpour[2*], MD; Andrew Bazemore[4], MD, MPH

[1]Department of Health Systems and Population Health Sciences, University of Houston, Houston, TX, United States

[2]Robert Graham Center for Policy Studies in Primary Care, Washington, DC, United States

[3]Centre for Health Informatics, Cumming School of Medicine, University of Calgary, Calgary, AB, Canada

[4]Center for Professionalism and Value in Health Care, American Board of Family Medicine, Washington, DC, United States

[*]these authors contributed equally

Corresponding Author:
Winston R Liaw, MD, MPH
Department of Health Systems and Population Health Sciences
University of Houston
4349 Martin Luther King Blvd
Houston, TX, 77204
United States
Phone: 1 7137439862
Email: winstonrliaw@gmail.com

Abstract

With conversational agents triaging symptoms, cameras aiding diagnoses, and remote sensors monitoring vital signs, the use of artificial intelligence (AI) outside of hospitals has the potential to improve health, according to a recently released report from the National Academy of Medicine. Despite this promise, the success of AI is not guaranteed, and stakeholders need to be involved with its development to ensure that the resulting tools can be easily used by clinicians, protect patient privacy, and enhance the value of the care delivered. A crucial stakeholder group missing from the conversation is primary care. As the nation's largest delivery platform, primary care will have a powerful impact on whether AI is adopted and subsequently exacerbates health disparities. To leverage these benefits, primary care needs to serve as a medical home for AI, broaden its teams and training, and build on government initiatives and funding.

KEYWORDS

Introduction

As noted in a 2019 report on artificial intelligence [1], Yuval Noah Harari wrote, "Humans were always far better at inventing tools than using them wisely" [2]. As data grow exponentially, this maxim is proving to be prescient in health care. From electronic health records (EHRs) and claims to smart devices, there are more electronic data than nucleotides in our individual DNA. Many note the potential of these data to advance the quintuple aim of better patient outcomes, population health, and health equity at lower costs while preserving clinician well-being. Although AI makes meaning from these gigabytes, it will fail without integration with the human and relational intelligence found in primary care.

A Landmark Report

The power of AI to advance health was highlighted in a recent National Academy of Medicine paper [3]. Its authors note that more affordable devices, broader internet access, and greater demand for digital health allow us to monitor health at home, augment telehealth, and predict which patients will get sick. This report will shape the AI conversation for years to come, and one of its contributions is a catalog of the uses of AI *outside* the hospital. Some examples are listed below:

- Conversational agents are triaging individuals with suspected COVID-19;

- Self-adaptive learning algorithms are working with continuous glucose monitors and insulin pumps to improve glucose control;
- Remote sensors are monitoring vital signs and transmitting data to cloud-based servers where they are acted upon, much like sensors, plugs, and appliances are powering smart homes.

Combining these data with text messages, social media, and geospatial coordinates permits the assessment of moods, outbreaks, and behavior changes. In the conclusion, the authors emphasize that questions on data standardization, usability, and reimbursement remain unanswered and go on to warn that AI can lead to privacy breaches and magnify biases, if diverse stakeholders are not engaged.

New Era, Same Mistakes

Despite its important insights, it is hard to ignore the absence of one stakeholder group—primary care—where *most* patients get *most* of their clinical care *most* of the time. Providing 50% of ambulatory visits and connecting with public health, primary care is vital to system transformation and needs to play a central role if AI is to enhance value. Prior efforts to integrate technology into health care neglected to engage primary care and resulted in systemic failures [4,5]. For example, family physicians are now spending more time with EHRs than patients, which has contributed to high rates of burnout [6]. Without engagement from primary care, AI could follow a similar path.

Primary care is important for multiple reasons. It is the largest delivery platform in the United States, accounting for 1 in 3 physicians [4,7]. Its presence is powerful enough to reduce mortality [8]. Its EHRs span organs and include behavioral and public health data, providing a comprehensive portrait of individual and population health. Family physicians, in particular, are distributed throughout the country, providing access to rural America [4]. Despite these benefits, only 5% of health care spending is devoted to primary care [9]. This misalignment has led to shortages and fragmentation. Like a wheel without a hub, care is not coordinated without primary care, and patients receive duplicate services and conflicting advice, contributing to greater waste [10,11].

AI has great potential to augment primary care and address these systemic challenges [12]. First, AI assistants can help with the ever-increasing demands for documentation within EHRs, a major factor driving burnout [13]. Using the same technology Alexa employs to turn on lights, play music, and order groceries, virtual assistants can transform speech into notes, and, in the future, can locate relevant information in the EHR, order labs, and adjust medications. By scanning the relevant primary care literature, AI can make recommendations so that patients receive care that is consistent with the current evidence. These innovations should allow primary care clinicians to spend less time locating and entering data and more time attending to patient relationships and solving their problems.

Second, AI can facilitate access to primary care. Conversational agents can interpret symptoms and assist with triage, helping patients to understand whether they need to be seen in the office

now, access emergency services, or monitor their symptoms at home. AI can combine this information with data from home devices such as internet-connected scales, glucometers, and, in the COVID-19 era, pulse oximeters to alert clinicians when patients need to be urgently seen. In this way, AI can serve as an early warning system to ensure that patients are evaluated at the right time and at the right place. Smartphones can analyze facial images, alert primary care clinicians when their patients' moods are deteriorating, and schedule visits before they get worse.

Third, AI can further enable a core feature of primary care—comprehensiveness. Video images can be used to diagnose diabetic retinopathy, dermatologic conditions, and Parkinson disease [14-16]. Applied appropriately, such applications could widen the scope of conditions retained in primary care and ensure that its clinicians are able to operate to the fullest extent of their training. Finally, AI can make care more person-centered. For instance, AI can use meal, geospatial, and activity tracking to provide the personalized health coaching needed to change behaviors and control chronic diseases. These applications would benefit from coordination with primary care so that coaching is reinforced during visits and informed by the patients' problems and medications. In addition to coaching, AI can predict the risk of acquiring a variety of diseases and identify the specific actions patients can take to mitigate the risk. Innovators in academia and industry are already using AI in these capacities, but more needs to be done to tailor these applications to primary care [12].

Lack of engagement with primary care in the development of these innovations creates a risk of limited implementation or adoption, and even worse, further fragmentation of health care delivery. Data niches could become more entrenched, with relevant information stored in separate locations. Patients could get conflicting information from sensors (eg, an alert indicating that a door is open but a video feed showing that it is closed) and may lack the knowledge to discern which signal to trust. Primary care is well suited to reconcile these conflicts. By eliciting preferences and values through shared decision-making, primary care clinicians help patients make sense of the data and coordinate with all team members (including patients) to refine treatment plans.

The Missing Link

To avoid additional failures, we propose three recommendations to ensure that primary care is involved in the future of AI.

1. In the Home and the Cloud

Primary care clinicians, informaticists, and researchers need to be involved in conversations regarding how health care data are collected, stored, or analyzed, with the primary care practice serving as the medical home for AI. Primary care can inform how data should be stored in the cloud, how algorithms ought to be used to adjust medications, and how tools are best integrated into primary care. This role is important for not only efficiency and effectiveness, but also equity. Because of its broad geographic distribution and focus on what patients need

within the context of their communities, primary care enhances equity, offsetting AI's tendency to widen disparities [17,18].

2. Transdisciplinary Teams

To fulfill this responsibility, health informaticists need to be integrated into primary care practices now. Similar to how AI will fail to adapt without primary care, primary care will fail to embrace AI without health informaticists. Team members with backgrounds in health informatics can connect the AI and primary care communities, by helping practices understand the benefits and limitations of AI, integrating AI into existing workflows, customizing AI applications for local contexts, and providing feedback from practices to AI developers. While these changes are daunting, primary care is skilled at adapting to the needs of its patients. As more and more struggled with mental health, primary care integrated behavioral health into their practices. As the number of medications grew, they successfully integrated pharmacists. Similarly, health informaticists can help practices and patients evolve in response to this digital revolution. In the long term, clinicians will need additional training in relevant disciplines, and researchers in other fields will need training in primary care. Ultimately, a transdisciplinary approach is needed to tailor AI to the complexity and longitudinality of primary care [19].

3. Government Facilitation

Achieving this vision will require governmental support for data standardization, funding, and governance. The Office of the National Coordinator for Health Information Technology can provide data standardization leadership. The Agency for Healthcare Research and Quality and the National Institutes of Health (NIH) can provide funding for primary care AI research. NIH funding supported a resource that provides data for thousands of patients who stayed in critical care units and that AI researchers use to develop AI tools [20]. A similar data set is needed for primary care. Given the time required to secure funding from the NIH, federal funders should coordinate with foundations and industry partners to stimulate activity in primary care AI now [21]. Finally, because AI needs to be trained and tested in multiple environments, a system that balances collaboration and data governance is needed. Federated learning, where algorithms from collaborators are tested locally, is one such approach, but more is needed to increase its adoption in primary care [22].

Time Flies Like an Arrow

AI has already transformed our cars, our homes, and our interactions. It has the power to positively impact care delivery, but also the potential to exacerbate existing health system failings. Assuring that AI is translated into knowledge will require engagement from all stakeholders. As a necessary component of AI, primary care is ready to assist.

References

1. Harari Y. 21 Lessons for the 21st Century. New York, NY: Random House; 2018.
2. Matheny M, Thadaney Israni S, Ahmed M, Whicher D, editors. Artificial Intelligence in Health Care: The Hope, the Hype, the Promise, the Peril. National Academy of Sciences. Washington, DC: National Academy of Medicine; 2019.
3. Aggarwal N, Ahmed M, Basu S, Curtin JJ, Evans BJ, Matheny ME, et al. Advancing Artificial Intelligence in Health Settings Outside the Hospital and Clinic. NAM Perspectives 2020 Nov 30. [doi: 10.31478/202011f]
4. Petterson S, McNellis R, Klink K, Meyers D, Bazemore A. The State of Primary Care in the United States. Robert Graham Center. 2018. URL: https://www.graham-center.org/content/dam/rgc/documents/publications-reports/reports/PrimaryCareChartbook.pdf [accessed 2022-03-04]
5. The Folsom Group. Communities of solution: the Folsom Report revisited. Ann Fam Med 2012 May 14;10(3):250-260 [FREE Full text] [doi: 10.1370/afm.1350] [Medline: 22585890]
6. Young RA, Burge SK, Kumar KA, Wilson JM, Ortiz DF. A Time-Motion Study of Primary Care Physicians' Work in the Electronic Health Record Era. Fam Med 2018 Feb;50(2):91-99 [FREE Full text] [doi: 10.22454/FamMed.2018.184803] [Medline: 29432623]
7. Johansen ME, Richardson CR. The Ecology of Medical Care Before and After the Affordable Care Act: Trends From 2002 to 2016. Ann Fam Med 2019 Nov 11;17(6):526-537 [FREE Full text] [doi: 10.1370/afm.2462] [Medline: 31712291]
8. Basu S, Berkowitz SA, Phillips RL, Bitton A, Landon BE, Phillips RS. Association of Primary Care Physician Supply With Population Mortality in the United States, 2005-2015. JAMA Intern Med 2019 Apr 01;179(4):506-514 [FREE Full text] [doi: 10.1001/jamainternmed.2018.7624] [Medline: 30776056]
9. Martin S, Phillips RL, Petterson S, Levin Z, Bazemore AW. Primary Care Spending in the United States, 2002-2016. JAMA Intern Med 2020 Jul 01;180(7):1019-1020 [FREE Full text] [doi: 10.1001/jamainternmed.2020.1360] [Medline: 32421142]
10. Osborn R, Moulds D, Squires D, Doty MM, Anderson C. International survey of older adults finds shortcomings in access, coordination, and patient-centered care. Health Aff (Millwood) 2014 Dec;33(12):2247-2255. [doi: 10.1377/hlthaff.2014.0947] [Medline: 25410260]

XSL·FO

RenderX

11.   Shrank WH, Rogstad TL, Parekh N. Waste in the US Health Care System: Estimated Costs and Potential for Savings.
       JAMA 2019 Oct 15;322(15):1501-1509. [doi: 10.1001/jama.2019.13978] [Medline: 31589283]
12.   Lin SY, Mahoney MR, Sinsky CA. Ten Ways Artificial Intelligence Will Transform Primary Care. J Gen Intern Med 2019
       Aug 14;34(8):1626-1630 [FREE Full text] [doi: 10.1007/s11606-019-05035-1] [Medline: 31090027]
13.   Peckham C. Medscape Lifestyle Report 2016: Bias and Burnout. Medscape. 2016 Jan 13. URL: https://www.medscape.com/
       slideshow/lifestyle-2016-overview-6007335 [accessed 2021-03-22]
14.   Skin Condition Questions? AI-Enabled Answers. Aysa. URL: https://askaysa.com/ [accessed 2021-03-22]
15.   Ke X. Tencent's AI Technology Assists Diagnosis of Parkinson's, Not to Replace Good Doctors. Tencent. 2019 Jul 11.
       URL: https://www.tencent.com/en-us/articles/2200927.html [accessed 2021-03-22]
16.   Verbraak FD, Abramoff MD, Bausch GCF, Klaver C, Nijpels G, Schlingemann RO, et al. Diagnostic Accuracy of a Device
       for the Automated Detection of Diabetic Retinopathy in a Primary Care Setting. Diabetes Care 2019 Apr;42(4):651-656.
       [doi: 10.2337/dc18-0148] [Medline: 30765436]
17.   Starfield B. Primary Care and Equity in Health: The Importance to Effectiveness and Equity of Responsiveness to Peoples'
       Needs. Humanity & Society 2009 Feb 01;33(1-2):56-73. [doi: 10.1177/016059760903300105]
18.   Obermeyer Z, Powers B, Vogeli C, Mullainathan S. Dissecting racial bias in an algorithm used to manage the health of
       populations. Science 2019 Oct 25;366(6464):447-453. [doi: 10.1126/science.aax2342] [Medline: 31649194]
19.   Liaw W, Kakadiaris IA. Primary Care Artificial Intelligence: A Branch Hiding in Plain Sight. Ann Fam Med 2020 May
       01;18(3):194-195 [FREE Full text] [doi: 10.1370/afm.2533] [Medline: 32393552]
20.   Medical Information Mart for Intensive Care. MIT-LCP. 2016 Sep. URL: https://mimic.physionet.org/about/mimic/ [accessed
       2021-01-21]
21.   Riley WT, Glasgow RE, Etheredge L, Abernethy AP. Rapid, responsive, relevant (R3) research: a call for a rapid learning
       health research enterprise. Clin Transl Med 2013 May 10;2(1):10 [FREE Full text] [doi: 10.1186/2001-1326-2-10] [Medline:
       23663660]
22.   Rieke N, Hancox J, Li W, Milletarì F, Roth HR, Albarqouni S, et al. The future of digital health with federated learning.
       NPJ Digit Med 2020 Sep 14;3(1):119 [FREE Full text] [doi: 10.1038/s41746-020-00323-1] [Medline: 33015372]

Abbreviations

AI: artificial intelligence
EHR: electronic health record
NIH: National Institutes of Health

XSL•FO
RenderX

<u>Viewpoint</u>

Information Extraction Framework for Disability Determination Using a Mental Functioning Use-Case

Ayah Zirikly[1,2,3], PhD; Bart Desmet[1], PhD; Denis Newman-Griffis[1,4], PhD; Elizabeth E Marfeo[1,5], MPH, PhD; Christine McDonough[1,6], PhD; Howard Goldman[1,7], MD, PhD; Leighton Chan[1], MPH, MD

[1]Rehabilitation Medicine Department, Clinical Center, National Institutes of Health, Bethesda, MD, United States

[2]Whiting School of Engineering, Johns Hopkins University, Baltimore, MD, United States

[3]Malone Center for Engineering in Healthcare, Johns Hopkins University, Baltimore, MD, United States

[4]Department of Biomedical Informatics, University of Pittsburgh, Pittsburgh, PA, United States

[5]Department of Occupational Therapy, Tufts University, Medford, MA, United States

[6]School of Health and Rehabilitation Science, University of Pittsburgh, Pittsburgh, PA, United States

[7]Department of Psychiatry, School of Medicine, University of Maryland, Baltimore, MD, United States

Corresponding Author:
Ayah Zirikly, PhD
Rehabilitation Medicine Department
Clinical Center
National Institutes of Health
10 Center Drive
Bethesda, MD, 20892
United States
Phone: 1 301 827 6558
Email: ayah.zirikly@nih.gov

Abstract

Natural language processing (NLP) in health care enables transformation of complex narrative information into high value products such as clinical decision support and adverse event monitoring in real time via the electronic health record (EHR). However, information technologies for mental health have consistently lagged because of the complexity of measuring and modeling mental health and illness. The use of NLP to support management of mental health conditions is a viable topic that has not been explored in depth. This paper provides a framework for the advanced application of NLP methods to identify, extract, and organize information on mental health and functioning to inform the decision-making process applied to assessing mental health. We present a use-case related to work disability, guided by the disability determination process of the US Social Security Administration (SSA). From this perspective, the following questions must be addressed about each problem that leads to a disability benefits claim: *When did the problem occur and how long has it existed? How severe is it? Does it affect the person's ability to work?* and *What is the source of the evidence about the problem?* Our framework includes 4 dimensions of medical information that are central to assessing disability—temporal sequence and duration, severity, context, and information source. We describe key aspects of each dimension and promising approaches for application in mental functioning. For example, to address temporality, a complete functional timeline must be created with all relevant aspects of functioning such as intermittence, persistence, and recurrence. Severity of mental health symptoms can be successfully identified and extracted on a 4-level ordinal scale from absent to severe. Some NLP work has been reported on the extraction of context for specific cases of wheelchair use in clinical settings. We discuss the links between the task of information source assessment and work on source attribution, coreference resolution, event extraction, and rule-based methods. Gaps were identified in NLP applications that directly applied to the framework and in existing relevant annotated data sets. We highlighted NLP methods with the potential for advanced application in the field of mental functioning. Findings of this work will inform the development of instruments for supporting SSA adjudicators in their disability determination process. The 4 dimensions of medical information may have relevance for a broad array of individuals and organizations responsible for assessing mental health function and ability. Further, our framework with 4 specific dimensions presents significant opportunity for the application of NLP in the realm of mental health and functioning beyond the SSA setting, and it may support the development of robust tools and methods for decision-making related to clinical care, program implementation, and other outcomes.

XSL•FO
RenderX

Introduction

Over the past 2 decades, the use of data-driven informatics techniques to aid in clinical decision-making has increased across the fields of computer science, bioinformatics, and medicine [1]. Natural language processing (NLP), which enables analysis of complex information recorded in narrative text format, has been a key driver of informatics successes in health care. Applications such as automated report analysis for clinical decision support and adverse event monitoring in the electronic health record (EHR) have been widely adopted [2-5]. However, informatics technologies for mental health have consistently lagged because of the complexity of measuring and modeling mental health and illness. The expansion of medical NLP technologies from clinical applications into the realm of complex administrative tasks such as claims evaluations and benefits administration [6,7] has highlighted the need for improved tools for analyzing information on mental health and function, which play a significant role in key health outcomes such as disability [8].

One of the primary goals of NLP in health data is to analyze narrative medical texts such as medical histories, physical examinations, and standardized assessments to extract the data needed to inform decision-making processes. The use of NLP to support these goals for management of mental health conditions has not yet been explored in depth. Our research group develops NLP models to support the information needs of the US Social Security Administration (SSA) disability determination process. Through its disability programs—Social Security Disability Insurance (SSDI) and Supplemental Security Income (SSI)—the SSA is the largest federal provider of financial assistance to workers with disabilities and their families. Because of the impact of functional abilities on both the individuals with disabilities and the society, it is essential that a person's functional abilities are characterized both comprehensively and efficiently in the disability determination process. Multiple sources of information are used to understand a person's ability to work. Given the complexity of the disability determination process, there is interest in developing approaches that enhance the validity of and confidence in the information across sources. While our work is motivated by the SSA's focused use-case, the SSA setting reflects fundamental challenges in the development and broad application of medical informatics technologies. The SSA leverages data from all types of health care providers and EHR systems across the United States. Therefore, informatics tools must be robust to significant heterogeneity in documentation—a known challenge for medical NLP research [9]. The volume of applications for disability benefits that the SSA must process is also extraordinarily high—over 2 million applications every year since 2004 [10]—and informatics tools must therefore support rapid

processing of high-volume data. Finally, and a key motivating factor for our work, the SSA's decision-making processes must incorporate diverse health and function information from all domains of human experience. The SSA setting thus provides an invaluable environment for learning how to translate the potential of NLP tools into practical, reliable tools for real-world applications.

Contributions of This Paper

In this paper, we propose a framework for the advanced application of NLP methods to identify, extract, and organize functioning information to inform the decision-making process applied to assessing functioning and disability. While the framework is applicable to mental and physical functioning use cases alike, this paper focuses on mental functioning. We found no literature that directly addresses our decision-support use-case for mental health and function; therefore, we developed a conceptual framework for synthesizing prior NLP research to create decision support tools for use in assessing SSA's definition of disability. Our framework includes 4 dimensions of medical information that are central to assessing disability—temporal sequence and duration, context, severity, and information source. Findings of this work are intended to inform the development of instruments that will support the decisions of disability adjudicators in the SSA's stepwise process of disability determination and have implications for a broad array of individuals and organizations responsible for assessing mental health function and ability. Further, our framework presents significant opportunity for the application of NLP in the realm of mental and physical health and functioning beyond the SSA setting, and it can support the development of robust tools and methods for decision-making related to clinical care, program implementation, and other outcomes.

Background

The US SSA administers the largest federal assistance programs in the United States, including 2 disability programs: SSDI and SSI. Both programs are based on a statutory definition of disability as the inability to engage in any substantial gainful activity by reason of any medically determinable physical or mental impairment(s), which can be expected to result in death or which has lasted or can be expected to last for a continuous period of not less than 12 months. The SSA's disability determination process is a stepwise process for evaluating individuals according to criteria that operationalize the statutory definition of disability. The process is based on federal regulatory standards that include both financial and medical criteria. Applicants are either allowed or denied at each step or move on to further evaluation in the subsequent steps. The process is administered by state Disability Determination Service agencies. In step 1, applicants are denied if they work and earn

more than the threshold for substantial gainful employment. In step 2, applicants are screened based on whether medical evidence supports the existence of a severe impairment. In step 3, the applicant's medical evidence is compared to codified clinical criteria for various medical impairments, called the Listing of Impairments (Listings). When impairments "meet" or "equal" the Listings, the applicants are allowed. Applicants who are not allowed at step 3 move on to steps 4 and 5, which assess vocational factors such as the "residual functional capacity" of the individual as well as the applicant's age, education, and relevant work experience. In step 4, adjudicators within the Disability Determination Service assess whether the applicant can work in any of their past jobs. If the adjudicator determines that an applicant can work in a previous job, the applicant is denied. Otherwise, in step 5, the Disability Determination Service adjudicators evaluate whether the applicant can perform any work in the national economy. There has been internal effort at the SSA to improve accuracy and timeliness of the disability adjudication process, and external groups have been engaged to assist with this. Expert panels and evaluations of the processes have resulted in recommendations for more systematic integration of functional information into adjudication decisions [11,12].

As part of the adjudication process, adjudicators amass a body of evidence referred to as the Medical Evidence of Record (MER), composed of information collected from multiple sources to characterize a person's potential ability to work. The MER forms the primary resource from which it is determined if an individual meets the SSA's statutory definition of disability. Therefore, the adjudicator must extract a variety of information from the MER, including medical evidence, medical opinion, and lay evidence, to support a decision on disability under this statutory definition.

A primary challenge for accuracy in the disability determination process is that adjudicators must access *all* relevant information from the MER for their decision, including information about both health conditions and functional abilities that relate to work. MER for a single individual may include dozens to hundreds of clinical reports, which imposes a significant burden on the adjudicator to rapidly process extensive medical evidence. Automated analysis of these documents with NLP thus has significant potential to assist adjudicators in the evidence review process and to support efficiency of the process. Our research group has developed novel NLP technologies for automated identification of functional status information in medical evidence [7], thus providing high-coverage retrieval of information related to mobility limitations [13-15] and categorization of this evidence according to the World Health Organization's International Classification of Functioning, Disability and Health [16]. Expansion of these technologies to mental health and function requires adaptation to the conceptual frameworks that characterize mental function, as outlined in the sections that follow.

For the purposes of this paper, we focus on mental health functioning, that is ways in which a person's underlying cognition, emotions, and behaviors affect their ability to perform daily activities including work tasks and participation, for example, a person's ability to regulate their emotions in stressful situations, multitask, or solve problems. This operationalization is based on the biopsychosocial model of health and function by the World Health Organization's International Classification of Functioning, Disability and Health. In this model, disability results from a gap between a person's underlying ability and the context in which they are performing various activities (eg, work participation [17,18]). This model highlights the fact that diagnostic information is necessary but not sufficient to understand a person's ability to participate in meaningful activities such as employment. In clinical contexts, information on functioning is critical to understanding the impact of conditions on people in their personal and environmental contexts and to develop an effective management plan. Recent work has demonstrated initial feasibility of applying NLP methods to mental health-related topics, including psychiatric readmission and symptoms of severe mental illness (SMI), as well as to mental health and suicide risk within nonclinical texts [19-23]. There is little evidence of the potential for NLP methods to characterize functional and behavioral manifestations of mental health in a person's daily life.

Information Needs for Analyzing Mental Health and Functioning

For disability adjudication, a wide array of information is needed about the following specific areas of mental functioning that a person uses in a work setting: understanding, remembering, or applying information; interacting with others; concentrating, persisting, or maintaining pace; and adapting and managing oneself. Evidence in the MER may reflect a physical or mental *impairment* that may affect these areas of functioning, an observed *limitation* in one of these areas of functioning, or both. The adjudicator's task is therefore to evaluate the level of severity or degree to which the medically determinable mental impairment affects the 4 areas of mental functioning (ie, limitations) and an individual's ability to function independently, appropriately, effectively, and on a sustained basis.

Thus, adjudicators must organize and synthesize the medical evidence in the following 4 distinct dimensions to understand the trajectory of mental function in an individual and its impact on work capacity: temporal information including sequence and duration, severity of extracted mentions of functioning, the context with respect to work and work-related information, and the source of the information.

We envision the use of NLP technologies to transform raw evidence found in medical records (MER in the SSA context) into a structured presentation of evidence illustrating each of these 4 factors to SSA adjudicators. Figure 1 illustrates the conceptual structure of such an analytic pipeline. Evidence is first extracted from the MER documents and then ordered into a temporal sequence, with each piece of evidence annotated with the severity of impact on function, the relevant work context, and the source of the evidence.

Figure 1. Conceptual pipeline for analysis of information on function in medical records. NLP: natural language processing.
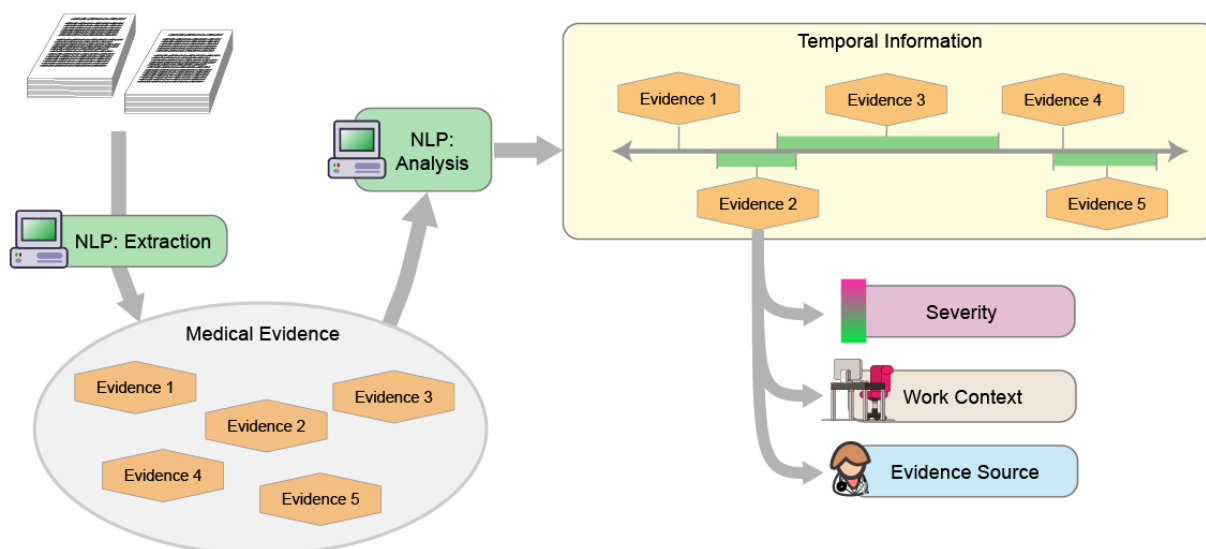
The remainder of this paper describes the existing NLP literature related to these 4 tasks and highlights how each can be addressed in the area of mental health and function.

Methods

Literature Search

We conducted a scoping review of NLP approaches, models, and methods to characterize functional status in free text in the biomedical, clinical, mental health, and disability domains from 1994 to 2021. We searched Google Scholar, which indexes not only PubMed but also conferences and workshops that may be relevant to our scope of interest such as the Association for Computational Linguistics (ACL) conferences, BioNLP, Clinical NLP, and CLPsych. Our search yielded a small number of publications in special workshops, such as AI4Function, that discuss the extraction of physical function information (eg, mobility) but do not address mental health function. We did not find any articles in our area of focus. To expand our search, we used keywords in Google Scholar related to the 4 dimensions of our proposed framework to find articles that describe approaches relevant to each dimension but an area different from functioning. Examples of keywords used include "temporal ordering," "clinical temporal ordering," "event extraction," "NLP and mental health," "symptom severity," "environmental context," "personal context," and "author attribution."

Findings From Existing Work

Disability in the SSA context is defined as the inability to engage in substantial gainful activity for at least 12 months because of a physical or mental impairment and is assessed both in terms of the *trajectory* of a person's function and the *context* of how it relates to work. As the disability adjudication process also involves collecting MER data from a variety of providers, it is critical to understand how different pieces of evidence relate to one another in terms of the perspective of the information's *source* (eg, the disability claimant, a medical professional with an established relationship with the claimant, an outside

consultant). Thus, for each piece of evidence in the MER, an adjudicator must be able to answer the following 4 questions: *When and for how long was an impairment or associated limitation true? How severe or intense is the impairment or limitation? Does the impairment or limitation affect work?* and *Who reported the impairment or limitation and how convincing is it as evidence?*

In this paper, we present an overview of relevant NLP research and methodologies that can help the adjudicator extract relevant information for these 4 questions. However, it is important to note that building solutions to address these 4 questions requires the ability to identify mental impairments either manually or via automated algorithms. In this paper, we choose to focus on addressing the 4 dimensions or questions only and assume that information on mental impairment is available. We justify our choice by the following factors: the availability of extraction systems that, given annotated data, can extract mentions of mental health impairments with high confidence in EHRs [24] and clinical text [21] and can extract these mentions using available International Classification of Diseases codes; and the novelty and urgency of the proposed 4 dimensions and the lack of available studies to address them. The mentions include observations such as "The patient was not able to concentrate on the given tasks for more than 5 minutes during the exam."

Temporal Information

Temporal information includes duration and temporal sequencing. In our use-case, the SSA's statutory definition of disability requires specific definition and sequential information. The disability is due to a mental impairment, so the impairment must precede the functional limitation.

Temporal sequencing or temporal reasoning has been an active research area in NLP and data mining for a long time. Its importance comes from its applicability to many tasks such as summarization [25], question answering [26], and medical informatics [27]. Given the similarity of the medical utilization task to our use-case, we will mainly focus on reviewing NLP

techniques developed in that area and how we can integrate them into our framework.

Temporal reasoning usually includes the following aspects: identifying the targeted events for the task (eg, treatments, diagnosis, symptoms, or medications); and defining time in a machine-readable way that is relevant to the domain and task and extracting temporal information related to the targeted events (eg, in a medical informatics setting, we care about the duration of symptoms or frequencies of medications) [28].

In the context of mental health functioning, the events of interest are events related to mental health conditions and impairment, which can be separated into the following 3 distinct categories: *persistent*, the impairment continues to exist over a prolonged time without interruptions of some criterion duration; *intermittent*, the impairment occurs at irregular intervals and is observed in a temporal sequence with interruptions greater than the criterion duration that defines persistent; and *recurring*, the impairment occurs periodically or repeatedly. Although we can think of recurring as a special case of intermittent, a recurrent mental functioning event is observed again after a period of some specified duration that is longer than the minimum duration defining intermittent.

While this list mainly focuses on the disability use-case, it presents a framework for researchers to structure their problem using all or a subset of our temporal formalization based on the targeted use-case, task, and domain. Our suggested framework differs from other NLP techniques for temporal sequencing because we need to consider nuances that accompany the mention of temporal information. For instance, a sentence such as "The patient reports having lack of interest mainly during the morning hours when it is the weekend," suggests the need for a system that can highlight the time: weekend morning hours, associated with lack of interest.

Although we introduce a slightly different framework for temporal sequencing, existing NLP methods can be applied to mental health functioning, especially given that most time expressions in medical notes are in the format of date and frequency (eg, how many times per week/day). For instance, temporal recognition and reasoning have played a significant role in information extraction systems [28-30]. Denny et al [29] developed a system that identifies the temporal information and status of colonoscopy events within EHRs with high precision and recall (>.9). In the area of mental health, Viani et al [28] focused on temporal expression extraction to help estimate the duration of untreated psychosis. The temporal information extraction helps in identifying in EHRs when the psychosis symptoms started (onset) and when the treatment was first initiated. Examples of temporal expressions from the paper include mentions such as "started hearing voices at the age of 16, these hallucinations were not elicited during today's exam." This is highly relevant to our use cases and to identify the 3 temporal formalizations of persistent, intermittent, and recurring.

To build NLP systems that can identify temporal information, the availability of annotated data sets with temporal information is critical. Although such data sets outside the clinical and medical domains have been publicly available and more easily accessible, such as the ACE 2005 Multilingual Training Corpus

[31], the clinical domain imposes more limitations, especially mental health, given the sensitivity of this information and privacy concerns. Examples of annotated data sets for temporal reasoning in clinical text are THYME corpus [32], where 1254 deidentified oncology notes from the Mayo Clinic have been annotated using the ISO-TimeML specification [33]. Sun et al [34] introduced one of the most popular data sets for temporal reasoning in their i2b2 data set that contains 310 discharge summaries. In both these data sets, the focus is on 4 time annotation categories: date, includes both actual dates in addition to mentions such as yesterday and tomorrow and duration, frequency, and time (eg, 3 PM, in the afternoon).

In another data genre, but within the area of mental health, there have been efforts to introduce temporally annotated data sets such as RSDD-Time [35]. This data set is extracted from social media posts that focus on self-reported patients who are diagnosed as having depression. The annotation includes temporal information relevant to when the diagnosis occurred and if the condition is still present.

Given our use-case, we believe that the i2b2 annotation scheme would serve our goals for identifying when the impairment or symptoms occurred and determining for how long the symptoms or impairments lasted.

With regard to methods, researchers used a variety of machine learning techniques such as logistic regression that is especially effective for a small training sample size of less than 500 [36]. Recent advances in the contextualized embeddings [37,38] improved the performance of NLP tasks, including temporal ordering in the clinical domain [39-41]. For instance, Med-BERT [42], a language model that is trained and fine-tuned on the EHR data set, yields a performance that is comparable to that of deep learning techniques on data sets that are almost 10 times larger.

Severity

The severity of a symptom or functional limitation is an important factor for psychological assessments and psychometric benchmarks, where it is often recorded using a 4-level ordinal scale or as a score that is discretized into that scale. A typical scale includes absent, mild, moderate, and severe labels [43]. The latter 3 labels are frequently employed for the disorders described in the Diagnostic and Statistical Manual of Mental Disorders Text Revision Fourth Edition (DSM-IV-TR), which permit severity specifiers. An advantage of this scale is that mental health clinicians and laypeople alike readily understand it, and it has been adopted in computational approaches for severity classification as well.

Filannino et al [44] describe an NLP shared task focused on symptom severity prediction in neuropsychiatric evaluation records with an exclusive focus on positive valence events, objects, or situations that are harmful but attractive to patients to the point that they are actively engaged despite the consequences. Positive valence is classified on the aforementioned 4-level scale at the patient level and assesses lifetime maximum severity. As such, this task differs from our approach in that it is not time dependent or resolved at the individual mention level. Filannino et al [44] report that in this

relaxed use-case, the task can be accomplished automatically with close to human performance.

Severity classification has also been actively researched in suicide risk assessment for patients and individuals on social media. For instance, Shing et al [45] and Zirikly et al [46] introduce an annotated Reddit data set for users with and without depression, each of whom received a suicide risk assessment score on a 4-level scale (none, low, moderate, high). Zirikly et al [46] organized a shared task for advanced automatic user suicide risk classification and provided baseline systems using deep learning models (eg, convolutional neural network) and machine learning models that require feature engineering. Examples of features that are commonly used in NLP methods for the mental health domain and emotion detection and classification are n-grams, lexicons such as the Linguistic Inquiry and Word Count [47], and emotion-word dictionaries [48], topic models, and Reddit usage metafeatures. Top-ranked systems could distinguish between low-risk and high-risk users, but fine-grained 4-level scale classification results indicate the need for further research.

Jackson et al [21] introduced an annotated data set for SMI using clinical text from the Clinical Record Interactive Search system in a cohort of 18,761 patients with SMI and 57,999 individuals without SMI. The authors used a support vector machine model to extract symptoms associated with SMI from discharge summaries. While the data and model for this task are relevant for the severity classification use-case, it does not address severity classification directly.

We can conclude that no severity classification models currently exist for mental health signs and symptoms, but there is a growing body of work on severity classification at the patient level. For clinical symptoms more broadly, Koleck et al [49] performed a systematic review of NLP approaches for processing symptoms in free-text EHR narratives. They found that out of 14 studies, the large majority used documentation occurrence or frequency of occurrence to investigate symptoms, and symptom severity was explicitly evaluated in only 1 study: Heintzelman et al [50] used NLP on oncology provider encounter notes to classify the severity of cancer patients' pain symptoms into no, some, controlled, or severe pain. Koleck et al [49] report accurate extraction of symptom severity with location and duration as important directions for future work on EHR NLP algorithms.

From our literature review, we find that for both mental and physical health, there is ample opportunity for novel work on severity classification of symptoms and functioning and for continued efforts at the patient level.

Context

The context in which a functional impairment or limitation is experienced or observed is critical to understanding its impact on work-related activities. Functional activity is an outcome of the interaction between an individual (including physical or cognitive impairments in addition to personal identities and preferences) and their physical, social, and cultural environment [51]. Characterization of this multidimensional relationship between environment, personal factors, and functioning is thus highly complex, as reflected by the wide variety of strategies used to capture contextual information in functional measurement [52]. Two themes have emerged in prior literature that make a useful distinction between different types of contextual factors: *social context* (ie, social determinants of health), broader characteristics of an individual's social situation such as socioeconomic status, education, zip code, and race and ethnicity identifiers, which inform available resources and opportunities for activity [53]; and *individual context*, factors that are more specific to an individual's activity performance, such as the physical environment for an activity, social roles such as work requirements, and personal preferences such as transportation access or personal values.

While research on social determinants of health has grown rapidly [54-57] due in part to their strong correlation with population-level health outcomes [58], research on individual context and environment—which more directly impacts functioning [59,60]—remains a significant challenge. Conceptual frameworks of disability have grown to recognize the role of both environmental factors and personal factors in functional outcomes, as seen in the World Health Organization's International Classification of Functioning, Disability and Health. Measures have been developed to characterize environmental factors of function, including physical [61] and psychosocial environment [62]. Such measures can be highly informative regarding functional outcomes [63]. However, they are not systematically used in clinical contexts [64,65] and some work-related aspects of environment remain underspecified even in conceptual models [66]. Functional assessment measures, on the other hand, frequently either control for environment (as in standardized performance measures) or embed environmental characteristics directly into the measurement of function [67,68] rather than capturing them as related variables. In either case, the details of a person's environment and its role in their functional outcomes are difficult to extract reliably.

Environmental factors are only one part of the contexts in which people function and must be combined with information on personal factors affecting functional outcomes. Two recent studies have developed steps toward systematically capturing personal values and capabilities to inform rehabilitative care for older adults, though automated analysis of this information remains a future direction [69,70]. Individual context is a largely unexplored area for NLP research due in part to the novelty of human functioning as an area of NLP application [71]. In an initial foray, Agaronnik et al [72] used NLP to capture wheelchair usage—which, as an assistive device, may be considered a contextual factor affecting functional outcomes—from clinical data and demonstrated clear utility of this information over structured billing and diagnosis codes alone. More broadly, the flexibility of free text and the availability of NLP tools to analyze it offers greater freedom for recording and analyzing information on salient contextual factors when the full power of more robust but burdensome environmental measures is not needed. We therefore highlight individual context, where social context meets individual activity, as a key direction for future NLP research to enable mental health and function analysis.

Source Attribution

In the context of the SSA, disability claims can include the following sources of information (Code of Federal Regulations, SSA): objective medical evidence, medical opinions, and lay evidence.

Objective medical evidence includes signs and laboratory tests reported by recognized medical sources. It is characterized by being quantifiable and discernable. This is highly important and indicative of the intensity and persistence of the symptoms and their impact (eg, how pain severity can affect work ability). Medical opinions include relevant information received from both medical and nonmedical sources. Examples of such information are daily activity and other factors relevant to functional limitations caused by pain or symptoms; location, duration, frequency, and intensity of pain or symptoms; and treatment or any other medication used to alleviate the pain or symptoms. Lay evidence consists of information outside of objective medial evidence or medical opinions—assessments of disability or functioning limitation provided by knowledgeable nonmedical sources such as family, teachers, social services personnel, and employers. This will complement the information provided in objective medical evidence and

medical opinions to better understand the impairments from multiple perspectives. Moreover, lay evidence is very insightful and important when medical evidence does not provide enough evidence for symptoms [73]. As we note, these types of evidence carry different levels of authority and support for the symptoms of the patient. Therefore, the adjudicators need to evaluate and address each evidence separately given its source to make a more comprehensive decision for disability eligibility. In this section, we will start with an overview of related work in NLP, followed by our recommendations to customize these efforts or build on them to address the needs to source attribution within our proposed framework.

Source attribution, as proposed, correlates with multiple similar tasks in NLP. We will start with an example to showcase options for NLP techniques we can adopt from: *The patient lacks interest in doing anything, his mom mentioned. When the doctor asked the patient, he claimed that he goes to work most of the time. At the end of the visit, the doctor diagnosed the patient with depression based on multiple assessments.* First, as we mentioned previously, we are assuming the availability of an extraction system that can identify and recognize mental health functioning and diagnoses statements. Table 1 depicts the mention, its source, and type of evidence.

Table 1. Examples of different mental health functioning.

Mention	Source	Type
The patient lacks interest in doing anything	Mother	Lay evidence (symptom)
He claimed that he goes to work most of the time	Patient	Medical opinion (daily activity)
The doctor diagnosed the patient with depression based on multiple assessments	Doctor	Objective medical evidence (diagnosis)

Identifying who made the statements is similar to the task of identifying author attribution in a dialogue or quoted speech [74]. Although regular expressions can capture simple cases of source attribution of impairments, such as "The patient said," Pareti et al [75] and O'Keefe et al [76] discussed more advanced techniques for quotes—direct and indirect—attribution in opinion mining. Although they show promising results, these methods have been geared and tested on newswire data. All these techniques require clean and well-structured data, an assumption that is hard to meet, especially given the noise presented in clinical notes [77].

There are some cases in which the doctor or medical expert omits mentioning the source in the note, especially when the observation is generated from them or another medical expert. In such cases, inferring source is difficult as it is not explicitly known or inferred (using coreference resolution). For these scenarios, we suggest that techniques be adopted from author attribution task. This task focuses on identifying the author of a text. This task has been well studied in multiple applications; the most traditional one is assigning anonymous literary to authors [78,79]. Additionally, it has been used in forensics to identify authors that are involved in internet-based activity in different text genres such as online messaging (eg, emails) [80], news text data set [81], and social media [82,83]. However, in our work, we focus on attribution of short text or sentences in notes.

Furthermore, we believe that it would be beneficial to adopt techniques from the intersection of author attribution and coreference resolution [84,85]. We see similarities with event extraction, where we focus on event attributes, mainly participants, when the event describes a mental health functioning mention. Techniques to extract multiple accounts from a narrative, such as the ones described by Zhang et al [86], can be adopted in our work to identify who made the observation or statement.

It is important to note that the attribution problem, as we propose it, requires systems that can identify mental health functioning mentions (eg, depression, lack of interest). For that goal, availability of annotated data to train and test machine learning systems is essential [87]. Although we earlier addressed the need for annotated data sets that have labels for mental health functioning, we need additional labels for the source attribution problem. The labels need to address who the source is and the type of evidence. However, it is worth pointing out that data sets that have labels solely on the level of evidence are sufficient for our targeted use-case. Labels can be assigned from the 3 types of evidence we mentioned above.

Discussion

Applications Beyond Disability Adjudication

We have presented a framework to support extraction of functional information in mental health, which includes 4 main

dimensions. There were no existing NLP applications that are directly applied to the characterization of temporality, severity, source, and context. However, we identified relevant work in mental health and other areas that could be used for the advanced application of NLP in the field. Temporal expression extraction and a relevant annotation scheme for identifying onset and duration were presented. A model for extracting severity of functional limitations was presented based on existing ordinal symptom severity ratings. An example of extraction of context was provided based on specific cases of wheelchair use in clinical settings. Finally, alternatives for source attribution were identified among existing approaches.

While our framework is tailored to the SSA disability benefits adjudication process, it has implications for a wide variety of applications outside the SSA context. For example, this approach could be used when extracting information for use in the medical case review process. This process requires expert review of patients' care history based on medical records to ensure that the treatment provided meets Medical Necessity Criteria. Additionally, this framework may be useful for other review activities, including informing the process for assessing eligibility determinations, individualized education programs, and educational placements for children under the Individuals with Disabilities Education Act. For this paper, we briefly highlight applications for mental health informatics research, functional assessment and program management in the health care setting, and consultations for case-based recommendations in treatment and managed care.

There is significant untapped potential for informatics technologies focusing on mental health and functioning, as evidenced by the interest of the mental health informatics research community in recent years. The use of informatics technologies has grown for the detection and diagnosis of mental health conditions [88], and the use as tools in mental health care delivery is beginning to be explored [89]. Our framework can inform the expansion of these technologies into a longer-term view of the trajectory of mental health and functioning in a person, thus improving the power of predictive analytics and presentation of health information to providers. Murnane et al [90] describe several technological needs for long-term mental health management, including incorporation of social contexts—a key component of our framework for NLP. Rigby et al [91] identified several aspects of mental health care that are still challenging for mental health informatics 2 decades later, including the importance of a longitudinal view. By identifying clear links to existing NLP research, our framework can serve to guide translational NLP research [92] in the mental health domain. This work can help identify both processes for translating existing NLP technologies into robust solutions for application in mental health research and care as well as new research questions for progress on the needs of mental health informatics.

There are numerous potential specific applications of our approach to using NLP for extraction of information from various sources to assist with the assessment of mental functioning. These are applications that require a review of medical and health-related information to assess functioning for the support of various clinical and other human service processes. The most common application would be reviews of clinical records to decide on a diagnosis or a course of treatment. A similar approach might be used by a managed care organization to determine the medical necessity of an episode of care or receipt of service. A consultant who is involved in the second opinion on a diagnosis or treatment plan could benefit from a decision-support tool that extracts all the information in a medical record that is relevant to mental functioning. Outside of health care settings, educational organizations and child welfare organizations might use such a clinical review to assess a student's need for special assistance or accommodation based on impairment in mental functioning. The development of an Individual Education Plan or a 504 plan [93] could use an NLP support tool to extract information from school and medical records to assess the need for special supports.

It is worth noting that this framework and its 4 key elements can be used for and generalized to any area of functioning within the SSA disability program and its statutory definition of work disability.

Support Tools for Disability Adjudication Need High Sensitivity

Current tools available to extract data related to mental health and function lack the level of sensitivity with respect to the elements in the MER on many types of mental functioning due to a mental impairment. While adjudicators ultimately need information on more fine-grained aspects of temporal sequencing using constructs such as intermittence, persistence, and recurrence, the main challenge is to create a complete timeline with all relevant aspects of functioning. Human decision makers assess the more fine-grained aspects of these characteristics of functioning. NLP systems thus need to extract the information without necessarily making fine-grained distinctions. One needs to know all the fine-grained elements to extract all relevant information even if the NLP tool does not need to make the distinctions. What is true for the granularity of temporality is also true for context, severity, and source, but most importantly, an NLP tool needs to be sensitive so that no information in the MER is overlooked.

Limitations and Challenges With Respect to NLP for Mental Health Function

Although the domain of mental health in general is attracting more NLP research, these studies focus on classification tasks in terms of diagnosis or identification of high-risk individuals and do not address how these impairments affect the patient's functioning in both personal and work environments. Thus, there will be challenges and obstacles as research evolves in this domain.

As we described in the paper, the domain of mental health in general and especially mental health functioning is ambiguous and highly semantic. This yields to different interpretations and inconsistencies in annotating documents with mental health functioning mentions and attributes, as consensus by humans is harder to attain. Furthermore, the lack of gold standard and manually annotated corpora for mental health functioning that are essential to build robust extraction solutions highlights the

need for the interested community to invest resources in building such corpora to further improve the performance of these solutions.

Although precision, specificity, and sensitivity (ie, recall) metrics are important, we believe that the interested entities, such as the SSA, can directly benefit from tools that focus on sensitivity (ie, higher recall) rather than higher precision and specificity. Such premise extends the invitation for research in categorization and relevance ranking to compensate for the low specificity and precision of such systems. Although we are aware of the importance of that line of research, this paper leaves this work for the future.

Conclusions and Future Vision

There is tremendous opportunity for the development and application of NLP tools and methods for the characterization of mental functioning. Although we found no literature that directly applied to the 4 main dimensions in the proposed framework, relevant tools and methods were identified. Research and development leveraging this existing work to tailor approaches for the extraction of temporality, severity, source, and context will yield substantial value to the use-case of disability determination and beyond. Future work should focus on developing relevant annotated data sets and tools trained on the key aspects of the 4 mental functioning domains.

Conflicts of Interest

None declared.

References

1. Ohno-Machado L. Realizing the full potential of electronic health records: the role of natural language processing. J Am Med Inform Assoc 2011 Sep 01;18(5):539-539 [FREE Full text] [doi: 10.1136/amiajnl-2011-000501] [Medline: 21846784]

2. Spyns P. Natural language processing in medicine: an overview. Methods Inf Med 1996 Dec;35(4-5):285-301. [Medline: 9019092]

3. Demner-Fushman D, Chapman WW, McDonald CJ. What can natural language processing do for clinical decision support? J Biomed Inform 2009 Oct;42(5):760-772 [FREE Full text] [doi: 10.1016/j.jbi.2009.08.007] [Medline: 19683066]

4. Young IJB, Luz S, Lone N. A systematic review of natural language processing for classification tasks in the field of incident reporting and adverse event analysis. Int J Med Inform 2019 Dec;132:103971. [doi: 10.1016/j.ijmedinf.2019.103971] [Medline: 31630063]

5. Mendonça EA, Haas J, Shagina L, Larson E, Friedman C. Extracting information on pneumonia in infants using natural language processing of radiology reports. J Biomed Inform 2005 Aug;38(4):314-321 [FREE Full text] [doi: 10.1016/j.jbi.2005.02.003] [Medline: 16084473]

6. Popowich F. Using text mining and natural language processing for health care claims processing. SIGKDD Explor Newsl 2005 Jun;7(1):59-66. [doi: 10.1145/1089815.1089824]

7. Desmet B, Porcino J, Zirikly A, Newman-Griffis D, Divita G, Rasch E. Development of Natural Language Processing Tools to Support Determination of Federal Disability Benefits in the U.S. In: Proceedings of the 1st Workshop on Language Technologies for Government and Public Administration (LT4Gov).: European Language Resources Association; 2020 Presented at: Language Resources and Evaluation Conference; 05/16/20; Marseille p. 1-6 URL: https://aclanthology.org/2020.lt4gov-1.1/

8. Courtney-Long EA, Carroll DD, Zhang QC, Stevens AC, Griffin-Blake S, Armour BS, et al. Prevalence of disability and disability type among adults--United States, 2013. MMWR Morb Mortal Wkly Rep 2015 Jul 31;64(29):777-783 [FREE Full text] [doi: 10.15585/mmwr.mm6429a2] [Medline: 26225475]

9. Carrell DS, Schoen RE, Leffler DA, Morris M, Rose S, Baer A, et al. Challenges in adapting existing clinical natural language processing systems to multiple, diverse health care settings. J Am Med Inform Assoc 2017 Sep 01;24(5):986-991 [FREE Full text] [doi: 10.1093/jamia/ocx039] [Medline: 28419261]

10. Social Security Administration. URL: https://www.ssa.gov/oact/STATS/table6c7.html [accessed 2022-02-18]

11. Stobo JD, McGeary M, Barnes DK, editors. Improving the social security disability decision process. Washington, DC: The National Academies Press; 2007.

12. Brandt DE, Houtenville AJ, Huynh MT, Chan L, Rasch EK. Connecting contemporary paradigms to the Social Security Administration's disability evaluation process. J Disabil Policy Stud 2011 Feb 07;22(2):116-128. [doi: 10.1177/1044207310396509]

13. Newman-Griffis D, Zirikly A. Embedding Transfer for Low-Resource Medical Named Entity Recognition: A Case Study on Patient Mobility. In: Proceedings of the BioNLP 2018 workshop.: Association for Computational Linguistics; 2018

Presented at: Annual Meeting of the Association for Computational Linguistics; 07/19/18; Melbourne, Australia p. 1-11 URL: https://aclanthology.org/W18-2301/ [doi: 10.18653/v1/w18-2301]

14. Newman-Griffis D, Fosler-Lussier E. HARE: a Flexible Highlighting Annotator for Ranking and Exploration. In: Proc Conf Empir Methods Nat Lang Process.: Association for Computational Linguistics; 2019 Presented at: Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP); 11/2019; Hong Kong, China p. 85-90 URL: https://aclanthology.org/D19-3015 [doi: 10.18653/v1/d19-3015]

15. Newman-Griffis D, Zirikly A, Divita G, Desmet B. Classifying the reported ability in clinical mobility descriptions. In: Proceedings of the 18th BioNLP Workshop and Shared Task.: Association for Computational Linguistics; 2019 Aug Presented at: Annual Meeting of the Association for Computational Linguistics; 07/2019; Florence, Italy p. 1-10 URL: https://aclanthology.org/W19-5001 [doi: 10.18653/v1/w19-5001]

16. Newman-Griffis D, Fosler-Lussier E. Automated coding of under-studied medical concept domains: linking physical activity reports to the International Classification of Functioning, Disability, and Health. Front Digit Health 2021 Mar 10;3:620828 [FREE Full text] [doi: 10.3389/fdgth.2021.620828] [Medline: 33791684]

17. Marfeo EE, Haley SM, Jette AM, Eisen SV, Ni P, Bogusz K, et al. Conceptual foundation for measures of physical function and behavioral health function for Social Security work disability evaluation. Arch Phys Med Rehabil 2013 Sep;94(9):1645-1652.e2 [FREE Full text] [doi: 10.1016/j.apmr.2013.03.015] [Medline: 23548543]

18. Chan F, Gelman J, Ditchman N, Kim JH, Chiu CY. The World Health Organization ICF model as a conceptual framework of disability. In: Understanding Psychosocial Adjustment to Chronic Illness and Disability: A Handbook for Evidence-Based Practitioners in Rehabilitation. NY: Springer; Jan 2009:23-49.

19. Rumshisky A, Ghassemi M, Naumann T, Szolovits P, Castro VM, McCoy TH, et al. Predicting early psychiatric readmission with natural language processing of narrative discharge summaries. Transl Psychiatry 2016 Oct 18;6(10):e921-e921 [FREE Full text] [doi: 10.1038/tp.2015.182] [Medline: 27754482]

20. Kjell ONE, Kjell K, Garcia D, Sikström S. Semantic measures: Using natural language processing to measure, differentiate, and describe psychological constructs. Psychol Methods 2019 Feb;24(1):92-115. [doi: 10.1037/met0000191] [Medline: 29963879]

21. Jackson RG, Patel R, Jayatilleke N, Kolliakou A, Ball M, Gorrell G, et al. Natural language processing to extract symptoms of severe mental illness from clinical text: the Clinical Record Interactive Search Comprehensive Data Extraction (CRIS-CODE) project. BMJ Open 2017 Jan 17;7(1):e012012 [FREE Full text] [doi: 10.1136/bmjopen-2016-012012] [Medline: 28096249]

22. Coppersmith G, Leary R, Crutchley P, Fine A. Natural language processing of social media as screening for suicide risk. Biomed Inform Insights 2018 Aug 27;10:1178222618792860 [FREE Full text] [doi: 10.1177/1178222618792860] [Medline: 30158822]

23. Calvo R, Milne D, Hussain M, Christensen H. Natural language processing in mental health applications using non-clinical texts. Nat Lang Eng 2017 Jan 30;23(5):649-685. [doi: 10.1017/s1351324916000383]

24. Wang Y, Wang L, Rastegar-Mojarad M, Moon S, Shen F, Afzal N, et al. Clinical information extraction applications: a literature review. J Biomed Inform 2018 Jan;77:34-49 [FREE Full text] [doi: 10.1016/j.jbi.2017.11.011] [Medline: 29162496]

25. Hirsch JS, Tanenbaum JS, Lipsky Gorman S, Liu C, Schmitz E, Hashorva D, et al. HARVEST, a longitudinal patient record summarizer. J Am Med Inform Assoc 2015 Mar;22(2):263-274 [FREE Full text] [doi: 10.1136/amiajnl-2014-002945] [Medline: 25352564]

26. Meng Y, Rumshisky A, Romanov A. Temporal Information Extraction for Question Answering Using Syntactic Dependencies in an LSTM-based Architecture. In: Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing.: Association for Computational Linguistics; 2017 Presented at: Conference on Empirical Methods in Natural Language Processing (EMNLP); 2017; Copenhagen, Denmark p. 887-896. [doi: 10.18653/v1/d17-1092]

27. Tourille J, Ferret O, Tannier X, Névéol A. Temporal information extraction from clinical text. In: Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics.: Association for Computational Linguistics; 2017 Presented at: Conference of the European Chapter of the Association for Computational Linguistics (EACL); 04/2017; Valencia, Spain p. 739-745. [doi: 10.18653/v1/e17-2117]

28. Viani N, Kam J, Yin L, Bittar A, Dutta R, Patel R, et al. Temporal information extraction from mental health records to identify duration of untreated psychosis. J Biomed Semantics 2020 Mar 10;11(1):2 [FREE Full text] [doi: 10.1186/s13326-020-00220-2] [Medline: 32156302]

29. Denny J, Peterson JF, Choma NN, Xu H, Miller RA, Bastarache L, et al. Extracting timing and status descriptors for colonoscopy testing from electronic medical records. J Am Med Inform Assoc 2010;17(4):383-388 [FREE Full text] [doi: 10.1136/jamia.2010.004804] [Medline: 20595304]

30. Moharasan G, Ho TB. Extraction of temporal information from clinical narratives. J Healthc Inform Res 2019 Feb 27;3(2):220-244. [doi: 10.1007/s41666-019-00049-0]

31. Doddington G, Mitchell A, Przybocki M, Ramshaw L, Strassel S, Weischedel R. The Automatic Content Extraction (ACE) Program - Tasks, Data, and Evaluation. In: Proceedings of the Fourth International Conference on Language Resources

and Evaluation (LREC).: European Language Resources Association (ELRA); 2004 Presented at: Language Resources and Evaluation Conference; 2004; Lisbon, Portugal URL: http://www.lrec-conf.org/proceedings/lrec2004/pdf/5.pdf

32. Styler WF, Bethard S, Finan S, Palmer M, Pradhan S, de Groen PC, et al. Temporal annotation in the clinical domain. Trans Assoc Comput Linguist 2014 Apr;2:143-154 [FREE Full text] [Medline: 29082229]

33. Pustejovsky J, Lee K, Bunt H, Romary L. ISO-TimeML: An International Standard for Semantic Annotation. In: Proceedings of the Seventh International Conference on Language Resources and Evaluation.: European Language Resources Association (ELRA); 2010 Presented at: Language Resources and Evaluation Conference; 05/2010; Valletta, Malta p. 394-397 URL: https://aclanthology.org/L10-1027/

34. Sun W, Rumshisky A, Uzuner O. Evaluating temporal relations in clinical text: 2012 i2b2 Challenge. J Am Med Inform Assoc 2013 Sep 01;20(5):806-813 [FREE Full text] [doi: 10.1136/amiajnl-2013-001628] [Medline: 23564629]

35. MacAvaney S, Desmet B, Cohan A, Soldaini L, Yates A, Zirikly A, et al. RSDD-Time: Temporal Annotation of Self-Reported Mental Health Diagnoses. In: Proceedings of the Fifth Workshop on Computational Linguistics and Clinical Psychology: From Keyboard to Clinic. 2018 Presented at: Conference of the North American Chapter of the Association for Computational Linguistics; 06/2018; New Orleans, LA p. 168-173 URL: https://aclanthology.org/W18-0618/ [doi: 10.18653/v1/w18-0618]

36. Rajkomar A, Oren E, Chen K, Dai AM, Hajaj N, Hardt M, et al. Scalable and accurate deep learning with electronic health records. NPJ Digit Med 2018 May 8;1(18):18 [FREE Full text] [doi: 10.1038/s41746-018-0029-1] [Medline: 31304302]

37. Devlin J, Chang MW, Lee K, Toutanova K. Bert: Pre-training of deep bidirectional transformers for language understanding. In: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies.: Association for Computational Linguistics; 2019 Presented at: Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies; 06/2019; Minneapolis, Minnesota p. 4171-4186. [doi: 10.18653/v1/N19-1423]

38. Peters ME, Neumann M, Iyyer M, Gardner M, Clark C, Lee K, et al. Deep contextualized word representations. In: Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies.: Association for Computational Linguistics; 2018 Presented at: Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies; 06/2018; New Orleans, LA p. 2227-2237 URL: https://aclanthology.org/N18-1202/ [doi: 10.18653/v1/N18-1202]

39. Lin C, Miller T, Dligach D, Sadeque F, Bethard S, Savova G. A BERT-based one-pass multi-task model for clinical temporal relation extraction. In: Proceedings of the 19th SIGBioMed Workshop on Biomedical Language Processing.: Association for Computational Linguistics; 2020 Presented at: Annual Meeting of the Association for Computational Linguistics; 07/2020; Online p. 70-75. [doi: 10.18653/v1/2020.bionlp-1.7]

40. Li Y, Rao S, Solares JRA, Hassaine A, Ramakrishnan R, Canoy D, et al. BEHRT: Transformer for Electronic Health Records. Sci Rep 2020 Apr 28;10(1):7155 [FREE Full text] [doi: 10.1038/s41598-020-62922-y] [Medline: 32346050]

41. Shang J, Ma T, Xiao C, Sun J. Pre-training of graph augmented transformers for medication recommendation. In: Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence. 2019 Presented at: International Joint Conference on Artificial Intelligence; 08/2019; Macao, China p. 5953-5959. [doi: 10.24963/ijcai.2019/825]

42. Rasmy L, Xiang Y, Xie Z, Tao C, Zhi D. Med-BERT: pretrained contextualized embeddings on large-scale structured electronic health records for disease prediction. NPJ Digit Med 2021 May 20;4(1):86 [FREE Full text] [doi: 10.1038/s41746-021-00455-y] [Medline: 34017034]

43. Spores JM. Clinician's Guide to Psychological Assessment and Testing: With Forms and Templates for Effective Practice. New York City: Springer Publishing Company; Sep 18, 2012:448.

44. Filannino M, Stubbs A, Uzuner Ö. Symptom severity prediction from neuropsychiatric clinical records: Overview of 2016 CEGS N-GRID shared tasks Track 2. J Biomed Inform 2017 Nov;75S:S62-S70 [FREE Full text] [doi: 10.1016/j.jbi.2017.04.017] [Medline: 28455151]

45. Shing H, Nair S, Zirikly A, Friedenberg M, Daumé III H, Resnik P. Expert, crowdsourced, and machine assessment of suicide risk via online postings. In: Proceedings of the Fifth Workshop on Computational Linguistics and Clinical Psychology: From Keyboard to Clinic.: Association for Computational Linguistics; 2018 Jun Presented at: Conference of the North American Chapter of the Association for Computational Linguistics; 2018; New Orleans, LA p. 25-36 URL: https://aclanthology.org/W18-0603/

46. Zirikly A, Resnik P, Uzuner Ö, Hollingshead K. CLPsych 2019 shared task: Predicting the degree of suicide risk in Reddit posts. In: Proceedings of the Sixth Workshop on Computational Linguistics and Clinical Psychology.: Association for Computational Linguistics; 2019 Jun Presented at: Conference of the North American Chapter of the Association for Computational Linguistics; 2019; Minneapolis, MN p. 24-33 URL: https://aclanthology.org/W19-3003/

47. Tausczik YR, Pennebaker JW. The psychological meaning of words: LIWC and computerized text analysis methods. J Lang Soc Psychol 2009 Dec 08;29(1):24-54. [doi: 10.1177/0261927x09351676]

48. Mohammad SM, Turney PD. Crowdsourcing a word-emotion association lexicon. Comput Intell 2013;29(3):436-465. [doi: 10.1111/j.1467-8640.2012.00460.x]

49. Koleck TA, Dreisbach C, Bourne PE, Bakken S. Natural language processing of symptoms documented in free-text narratives of electronic health records: a systematic review. J Am Med Inform Assoc 2019 Apr 01;26(4):364-379 [FREE Full text] [doi: 10.1093/jamia/ocy173] [Medline: 30726935]

50. Heintzelman NH, Taylor RJ, Simonsen L, Lustig R, Anderko D, Haythornthwaite JA, et al. Longitudinal analysis of pain in patients with metastatic prostate cancer using natural language processing of medical record text. J Am Med Inform Assoc 2013 Sep 01;20(5):898-905 [FREE Full text] [doi: 10.1136/amiajnl-2012-001076] [Medline: 23144336]

51. Playford D. The International Classification of Functioning, Disability, and Health. In: Dietz V, Ward N, editors. Oxford Textbook of Neurorehabilitation. Oxford, UK: Oxford University Press; Feb 2015:3-7.

52. Iezzoni LI, Marsella SA, Lopinsky T, Heaphy D, Warsett KS. Do prominent quality measurement surveys capture the concerns of persons with disability? Disabil Health J 2017 Apr;10(2):222-230. [doi: 10.1016/j.dhjo.2017.01.007] [Medline: 28185857]

53. World Health Organization. In: Wilkinson R, Marmot M, editors. Social Determinants of Health: The Solid Facts 2nd ed. Copenhagen: WHO Regional Office for Europe; 2003.

54. Conway M, Keyhani S, Christensen L, South BR, Vali M, Walter LC, et al. Moonstone: a novel natural language processing system for inferring social risk from clinical narratives. J Biomed Semantics 2019 Apr 11;10(1):6 [FREE Full text] [doi: 10.1186/s13326-019-0198-0] [Medline: 30975223]

55. Dorr D, Bejan CA, Pizzimenti C, Singh S, Storer M, Quinones A. Identifying patients with significant problems related to social determinants of health with natural language processing. Stud Health Technol Inform 2019 Aug 21;264:1456-1457. [doi: 10.3233/SHTI190482] [Medline: 31438179]

56. Deferio JJ, Breitinger S, Khullar D, Sheth A, Pathak J. Social determinants of health in mental health care and research: a case for greater inclusion. J Am Med Inform Assoc 2019 Aug 01;26(8-9):895-899 [FREE Full text] [doi: 10.1093/jamia/ocz049] [Medline: 31329877]

57. Feller DJ, Bear Don't Walk Iv OJ, Zucker J, Yin MT, Gordon P, Elhadad N. Detecting social and behavioral determinants of health with structured and free-text clinical data. Appl Clin Inform 2020 Jan 04;11(1):172-181 [FREE Full text] [doi: 10.1055/s-0040-1702214] [Medline: 32131117]

58. Closing the gap in a generation: health equity through action on the social determinants of health - Final report of the commission on social determinants of health. World Health Organization. 2008. URL: https://www.who.int/publications/i/item/WHO-IER-CSDH-08.1 [accessed 2022-02-22]

59. Stolwijk C, Castillo-Ortiz JD, Gignac M, Luime J, Boonen A, OMERACT Worker Productivity Group. Importance of contextual factors when measuring work outcome in ankylosing spondylitis: a systematic review by the OMERACT Worker Productivity Group. Arthritis Care Res (Hoboken) 2015 Sep 26;67(9):1316-1327 [FREE Full text] [doi: 10.1002/acr.22573] [Medline: 25732705]

60. Sinclair CM, Meredith P, Strong J, Feeney R. Personal and contextual factors affecting the functional ability of children and adolescents with chronic pain: a systematic review. J Dev Behav Pediatr 2016 May;37(4):327-342. [doi: 10.1097/DBP.0000000000000300] [Medline: 27096569]

61. Brownson RC, Hoehner CM, Day K, Forsyth A, Sallis JF. Measuring the built environment for physical activity: state of the science. Am J Prev Med 2009 Apr;36(4 Suppl):S99-123.e12 [FREE Full text] [doi: 10.1016/j.amepre.2009.01.005] [Medline: 19285216]

62. Stansfeld S, Candy B. Psychosocial work environment and mental health--a meta-analytic review. Scand J Work Environ Health 2006 Dec;32(6):443-462 [FREE Full text] [doi: 10.5271/sjweh.1050] [Medline: 17173201]

63. Orstad SL, McDonough MH, Stapleton S, Altincekic C, Troped PJ. A systematic review of agreement between perceived and objective neighborhood environment measures and associations with physical activity outcomes. Environ Behav 2016 Sep 29;49(8):904-932. [doi: 10.1177/0013916516670982]

64. Madans J. Proposed purpose of an internationally comparable general disability measure. 2004 Presented at: Third Meeting Washington Group on Disability Statistics; February 19-20, 2004; Brussels, Belgium.

65. Altman BM. Appendix A: Population Survey Measures of Functioning: Strengths and Weaknesses. In: Wunderlich GS, editor. Improving the Measurement of Late-Life Disability in Population Surveys: Beyond ADLs and IADLs: Summary of a Workshop National Academies of Sciences, Engineering, and Medicine. 2009. Improving the Measurement of Late-Life Disability in Population Surveys: Beyond ADLs and IADLs: Summary of a Workshop. Washington, DC: The National Academies Press; 2009:99-156.

66. Heerkens YF, de Brouwer CP, Engels JA, van der Gulden JW, Kant I. Elaboration of the contextual factors of the ICF for Occupational Health Care. Work 2017;57(2):187-204. [doi: 10.3233/WOR-172546] [Medline: 28582939]

67. Haley SM, Coster WJ, Binda-Sundberg K. Measuring physical disablement: the contextual challenge. Phys Ther 1994 May;74(5):443-451. [doi: 10.1093/ptj/74.5.443] [Medline: 8171106]

68. Jette DU, Halbert J, Iverson C, Miceli E, Shah P. Use of standardized outcome measures in physical therapist practice: perceptions and applications. Phys Ther 2009 Feb;89(2):125-135. [doi: 10.2522/ptj.20080234] [Medline: 19074618]

69. Stephens C, Breheny M, Mansvelt J. Healthy ageing from the perspective of older people: a capability approach to resilience. Psychol Health 2015 Apr 29;30(6):715-731. [doi: 10.1080/08870446.2014.904862] [Medline: 24678916]

70. Yeung P, Breheny M. Quality of life among older people with a disability: the role of purpose in life and capabilities. Disabil Rehabil 2021 Jan;43(2):181-191. [doi: 10.1080/09638288.2019.1620875] [Medline: 31335217]

71. Newman-Griffis D, Porcino J, Zirikly A, Thieu T, Camacho Maldonado J, Ho P, et al. Broadening horizons: the case for capturing function and the role of health informatics in its use. BMC Public Health 2019 Oct 15;19(1):1288 [FREE Full text] [doi: 10.1186/s12889-019-7630-3] [Medline: 31615472]

72. Agaronnik ND, Lindvall C, El-Jawahri A, He W, Iezzoni LI. Challenges of Developing a Natural Language Processing Method With Electronic Health Records to Identify Persons With Chronic Mobility Disability. Arch Phys Med Rehabil 2020 Oct;101(10):1739-1746 [FREE Full text] [doi: 10.1016/j.apmr.2020.04.024] [Medline: 32446905]

73. Committee on Psychological Testing, Including Validity Testing, for Social Security Administration Disability Determinations, Board on the Health of Select Populations; Institute of Medicine. Psychological Testing in the Service of Disability Determination. Washington DC: National Academies Press; Jun 29, 2015.

74. Elson D, McKeown K. Automatic attribution of quoted speech in literary narrative. In: AAAI. 2010 Presented at: The Twenty-Fourth AAAI Conference on Artificial Intelligence; July 2010; Atlanta, Georgia.

75. Pareti S, O'Keefe T, Konstas I. Automatically Detecting and Attributing Indirect Quotations. 2013 Presented at: Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing,; October 18-21, 2013; Seattle, Washington URL: https://openreview.net

76. O'Keefe T. A sequence labelling approach to quote attribution. 2012 Jul Presented at: EMNLP-CoNLL '12: Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning; July 2012; Jeju Island, Korea.

77. Nguyen H, Patrick J. Text Mining in Clinical Domain: Dealing with Noise. 2016 Presented at: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining; August 13, 2016; San Francisco, CA. [doi: 10.1145/2939672.2939720]

78. Burrows J. 'Delta': a measure of stylistic difference and a guide to likely authorship. Lit Linguistics Comput 2002 Sep 01;17(3):267-287. [doi: 10.1093/llc/17.3.267]

79. Hoover DL. Testing Burrows's Delta. Lit Linguistics Comput 2004 Nov 01;19(4):453-475. [doi: 10.1093/llc/19.4.453]

80. Nirkhi S, Dharaskar R, Thakare V. Authorship verification of online messages for forensic investigation. Procedia Comput Sci 2016;78:640-645. [doi: 10.1016/j.procs.2016.02.111]

81. Lambers M, Veenman CJ. Forensic Authorship Attribution Using Compression Distances to Prototypes. 2009 Presented at: International Workshop on Computational Forensics; August 13-14, 2009; The Hague, The Netherlands p. 13-24. [doi: 10.1007/978-3-642-03521-0_2]

82. Rocha A, Scheirer WJ, Forstall CW, Cavalcante T, Theophilo A, Shen B, et al. Authorship attribution for social media forensics. IEEE Trans Inform Forensic Secur 2017 Jan;12(1):5-33. [doi: 10.1109/tifs.2016.2603960]

83. Frye RH, Wilson DC. Defining Forensic Authorship Attribution for Limited Samples from Social Media. 2018 Presented at: FLAIRS Conference 2018; September 10-14, 2018; Melbourne Beach.

84. O'Keefe T. Examining the Impact of Coreference Resolution on Quote Attribution. 2013 Presented at: Proceedings of Australasian Language Technology Association Workshop; 2013; Brisbane, Australia p. 43-52.

85. Almeida M. A Joint Model for Quotation Attribution and Coreference Resolution. 2014 Presented at: The 14th Conference of the European Chapter of the Association of Computational Linguistics; April 2014; Gothenburg, Sweden p. 39-48. [doi: 10.3115/v1/e14-1005]

86. Zhang H, Boons F, Batista-Navarro R. Whose story is it anyway? Automatic extraction of accounts from news articles. Inf Process Manag 2019 Sep;56(5):1837-1848. [doi: 10.1016/j.ipm.2019.02.012]

87. Walker V. The Need for Annotated Corpora from Legal Documents, and for (Human) Protocols for Creating Them: The Attribution Problem. Maurice A. Deane School of Law at Hofstra University. Scholarly Commons at Hofstra Law. 2016. URL: https://scholarlycommons.law.hofstra.edu/cgi/viewcontent.cgi?article=2231&context=faculty_scholarship [accessed 2022-02-22]

88. Shatte ABR, Hutchinson DM, Teague SJ. Machine learning in mental health: a scoping review of methods and applications. Psychol Med 2019 Jul;49(9):1426-1448. [doi: 10.1017/S0033291719000151] [Medline: 30744717]

89. Kemp J, Zhang T, Inglis F, Wiljer D, Sockalingam S, Crawford A, et al. Delivery of compassionate mental health care in a digital technology-driven age: scoping review. J Med Internet Res 2020 Mar 06;22(3):e16263 [FREE Full text] [doi: 10.2196/16263] [Medline: 32141833]

90. Murnane EL, Walker TG, Tench B, Voida S, Snyder J. Personal informatics in interpersonal contexts. Proc ACM Hum-Comput Interact 2018 Nov;2(CSCW):1-27. [doi: 10.1145/3274396]

91. Rigby M, Lindmark J, Furlan PM. The importance of developing an informatics framework for mental health. Health Policy 1998 Jul;45(1):57-67. [doi: 10.1016/s0168-8510(98)00028-1] [Medline: 10183013]

92. Newman-Griffis D. Translational NLP: A New Paradigm and General Principles for Natural Language Processing Research. 2021 Jun Presented at: Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics; June 2021; Online p. 4125-4138. [doi: 10.18653/v1/2021.naacl-main.325]

93. Bishop T. Mental disorders and learning disabilities in children and adolescents: learning disabilities. FP Essent 2018 Dec;475:18-22. [Medline: 30556687]

XSL•FO
RenderX

Abbreviations

ACL: Association for Computational Linguistics
DSM-IV-TR: Diagnostic and Statistical Manual of Mental Disorders Text Revision Fourth Edition
EHR: electronic health record
MER: Medical Evidence of Record
NLP: natural language processing
SMI: severe mental illness
SSA: Social Security Administration
SSDI: Social Security Disability Insurance
SSI: Supplemental Security Income

XSL•FO
RenderX

Original Paper

Deriving Weight From Big Data: Comparison of Body Weight Measurement–Cleaning Algorithms

Richard Evans[1], MS; Jennifer Burns[1], MHSA; Laura Damschroder[1], MS, MPH; Ann Annis[1,2], RN, PhD; Michelle B Freitag[1], MPH; Susan Raffa[3,4], PhD; Wyndy Wiitala[1], PhD

[1]Center for Clinical Management Research, Veterans Health Administration, Ann Arbor, MI, United States

[2]College of Nursing, Michigan State University, Lansing, MI, United States

[3]National Center for Health Promotion and Disease Prevention, Veterans Health Administration, Durham, NC, United States

[4]Department of Psychiatry & Behavioral Sciences, Duke University School of Medicine, Durham, NC, United States

Corresponding Author:
Richard Evans, MS
Center for Clinical Management Research
Veterans Health Administration
2215 Fuller Road
Mail Stop 152
Ann Arbor, MI, 48105
United States
Phone: 1 248 910 3441
Email: Richard.Evans8@va.gov

Abstract

Background: Patient body weight is a frequently used measure in biomedical studies, yet there are no standard methods for processing and cleaning weight data. Conflicting documentation on constructing body weight measurements presents challenges for research and program evaluation.

Objective: In this study, we aim to describe and compare methods for extracting and cleaning weight data from electronic health record databases to develop guidelines for standardized approaches that promote reproducibility.

Methods: We conducted a systematic review of studies published from 2008 to 2018 that used Veterans Health Administration electronic health record weight data and documented the algorithms for constructing patient weight. We applied these algorithms to a cohort of veterans with at least one primary care visit in 2016. The resulting weight measures were compared at the patient and site levels.

Results: We identified 496 studies and included 62 (12.5%) that used weight as an outcome. Approximately 48% (27/62) included a replicable algorithm. Algorithms varied from cutoffs of implausible weights to complex models using measures within patients over time. We found differences in the number of weight values after applying the algorithms (71,961/1,175,995, 6.12% to 1,175,177/1,175,995, 99.93% of raw data) but little difference in average weights across methods (93.3, SD 21.0 kg to 94.8, SD 21.8 kg). The percentage of patients with at least 5% weight loss over 1 year ranged from 9.37% (4933/52,642) to 13.99% (3355/23,987).

Conclusions: Contrasting algorithms provide similar results and, in some cases, the results are not different from using raw, unprocessed data despite algorithm complexity. Studies using point estimates of weight may benefit from a simple cleaning rule based on cutoffs of implausible values; however, research questions involving weight trajectories and other, more complex scenarios may benefit from a more nuanced algorithm that considers all available weight data.

XSL•FO
RenderX

Introduction

Background

The use of electronic health records (EHRs) by health care systems has rapidly increased during the last 2 decades [1], making vast amounts of clinical information available for use in research and evaluation efforts [2,3]. However, there are issues associated with using EHR data, including a lack of control over data definitions and data collection processes [4] as well as methodological challenges associated with processing and transforming raw, messy EHR [5] data into research-ready data that can be meaningfully used for research and evaluation [6]. For these reasons, many have called for increased transparency regarding data cleaning efforts, methods to assess EHR data quality [7], and increased reporting and sharing of methods for selecting clinical codes [8,9].

Obesity is associated with increased risk of a wide range of medical problems, including diabetes, hypertension, high blood cholesterol, cardiovascular events, bone and joint problems, and sleep apnea [10]. Clinicians frequently advise patients to lose weight to help prevent or delay the onset of chronic disease [11]. Accordingly, obesity is a major public health challenge for the United States; compared with patients of normal weight, patients with obesity have higher inpatient costs, more outpatient visits and costs, and more spending on prescription drugs [12]. Thus, patient weight represents a frequently used measure for many researchers and evaluators. It may be included as a risk factor in studies seeking to predict adverse medical events, as a covariate in studies that seek to adjust for the effect of baseline weight when examining the association between another variable (eg, treatment) and an outcome, or as an outcome in studies examining the effects of a measure (eg, intervention) on patient weight or weight change over time.

Despite being a common clinical measure, there is no standard for processing and cleaning EHR weight data for use in research and evaluation studies. Researchers are often left to select and replicate a method described by others or develop their own algorithms to define weight measures for analyses, resulting in many different definitions in the published literature [13]. These definitions range from simple cutoffs for implausible values to more computationally complex algorithms requiring significant coding and processing capacity, as well as difficulties in replicating for other studies. Furthermore, it is unknown how resulting weight measures may vary based on how researchers process and clean the data; subsequently, the impact of algorithm choice on results and research findings is also unknown.

Study Objective

The objectives of this study include (1) comparing algorithms for extracting and processing clinical weight measures from EHR databases and (2) providing recommendations for the use of algorithms. We used measures of patient weight from the Corporate Data Warehouse (CDW) of the Veterans Health Administration (VHA) to accomplish these objectives. The VHA includes a network of medical centers that rely on a system-wide integrated EHR system. Patient data are extracted from EHR records nightly and uploaded to a centralized CDW, which comprises relational data tables that can be accessed by data analysts, including researchers. Users extract data from the CDW and typically perform simple data checks to verify accuracy. More complex algorithms may be used, especially in research; for example, to ensure that the amount of missing data does not exceed a prespecified threshold [14].

Methods

Cohort and Data Sources

We included cohorts of VHA patients based on two calendar year periods: 2008 and 2016. Previous work suggests that data quality for some CDW data fields has improved over time in terms of cleanliness and data capture [15,16]. Thus, selecting 2 time points allowed us to compare the quality and quantity of data between these time points. For each year, we randomly sampled 100,000 patients aged ≥18 years with at least one primary care visit (VHA Stop Code 323) during the cohort year, with the first primary care visit serving as the index date. There were no restrictions on facility or region; thus, our cohorts represent a national sample. We excluded patients with any International Classification of Diseases, 9th or 10th revision codes, or Current Procedural Terminology codes for pregnancy within 2 years before and 2 years after the index date, which we henceforth refer to as the *collection period*. Our detailed approach is described in Multimedia Appendix 1 [17-28].

We collected all weight and height measurements from the CDW vital sign table during the collection period. If a patient had more than one height measurement during the 4 years, we used the modal value to determine a single measure of height for each patient. In the event that an individual only had 2 recorded height values, the last value was chosen when height was arranged in ascending order by collection date. We calculated BMI by dividing weight in kilograms by height in meters squared. All weight and height data were cleansed of any nonnumeric characters, converting commas to decimals where appropriate.

Weight-Cleaning Algorithms

Previously, our team conducted a systematic literature review to identify studies that used patient weight outcome measures from the VHA CDW [13]. We identified 39 published studies that used the CDW to define patient weight outcomes. Of the 39 studies, 33 (85%) [17-49] included a weight-cleaning algorithm that could be implemented and replicated in this study. In this paper, we present 12 algorithms [17-28] representing the breadth of methods used in cleaning body weight measurements and provide details about the remaining algorithms in Multimedia Appendix 1 and in our GitHub repository [50].

For comparison, we divided the 12 algorithms into two conceptual groups: (1) those that included all weight measurements during a specified time frame and (2) those that were period-specific. A brief description of the key differences between algorithms by group is shown in Table 1. Period-specific algorithms were those that selected *baseline, 6-month, and 12-month* periods and included weight measurements during specified windows around those dates. Note that not all algorithms fit exactly into these groups. For instance, we classified the algorithm used in the study by Noël

et al [27] as a *period-specific* algorithm, as it is based on fiscal quarters but uses all data within each quarter to define median weights. Similarly, the algorithm by Jackson et al [21] involves taking the arithmetic mean of all weight measurements collected between arbitrarily chosen time points.

All algorithms were recreated from the methods sections described in the relevant publications and translated into pseudocode and then into R (version 3.6.1; R Foundation for Statistical Computing) or SAS (version 9.4; SAS Institute) code (Multimedia Appendix 1, section 2, and web-based supplemental materials [50]).

Table 1. Conceptual description of main exclusions after applying each algorithm[a].

Conceptual group	Exclusions based on algorithm
All weight measures	
Buta et al [18]	• Patients with ≤1 weight value • BMI <11 or >70
Chan and Raffa [19]	• Weights <23 kg or >340 kg • Weights >3 SD from mean
Maguen et al [26]	• Weights <32 kg or >318 kg • Weights where the absolute value of conditional residual from linear mixed model ≥10
Breland et al [17]	• Weights <34 kg or >318 kg • Weight values that fell outside of specific ratios calculated within patients over time
Maciejewski et al [25]	• Weight values associated with large SDs calculated on a rolling basis
Littman et al [24]	• Weights <34 kg or >272 kg • Weights where difference from mean >SD • Weights where SD was >10% of the mean
Period-specific	
Rosenberger et al [28]	• Patients with <K number of weight measures; K chosen by researcher • Weights outside of 6-month time points
Noël et al [27]	• Weights ≤32 kg or ≥318 kg • Patients with too few values to compute median within fiscal quarters
Kazerooni and Lim [23]	• Weights outside of windows around 3 periods • Patients missing data in any of the 3 periods
Jackson et al [21]	• Weights <34 kg or >318 kg • Weights outside of 90-day window of each time point
Goodrich et al [20]	• Weights <36 kg or >227 kg • Patients with >45 kg change between periods (baseline and 6 and 12 months) • Weights outside of 30-day window of each time point
Janney et al [22]	• Weights <41 kg or >272 kg at baseline • Weights outside of 30-day window of baseline and 60-day window of 6- and 12-month period • Weights resulting in >45 kg change during study

[a]Details of each algorithm, including code, excerpts from published methods, and pseudocode, can be found in Multimedia Appendix 1, section 2, and the project GitHub [50].

Methods to Compare Algorithms

Descriptive Statistics

All algorithms were applied to the data for both cohorts and compared based on descriptive statistics, including the number of weight measures and patients retained and the mean, SD, median, and range of weight values. For comparison, we also included descriptive statistics based on the raw, unprocessed weight data during the study time frame.

Weight as a Predictor

Weight is often used as a risk factor or covariate in statistical models to predict health outcomes. We present an example showing the association between baseline weight and *new-onset* diabetes to compare algorithms in this context. For this analysis, we excluded patients with diabetes before the study index date and we defined new-onset diabetes as the presence of 2 or more diabetes diagnosis codes after the patient's index date. To create baseline weight measures for each patient, all 12 algorithms were first applied to each cohort, then weight measurements

were collected given a 60-day window on or before the index date (ie, 30 days before to 30 days after the index date). The resulting baseline weight measure was the measurement that occurred on the closest day to the index date after cleaning the weight data. We then used 13 distinct logistic regression models to obtain odds ratios (ORs) for the effect of patient weight on new-onset diabetes.

Weight Change

A common metric used in weight loss evaluation studies involves *weight loss ≥5%*, where weight change is assessed over a 1-year period [11]. We applied each algorithm to our cohorts to compare algorithms on this metric. After cleaning the weight data, we used a 60-day window to define initial weight values and included the weight measurement taken on the closest day to the index date. To define the 1-year follow-up weight, we again used a 60-day window around the date 1 year after the baseline, keeping the closest weight measurement. In addition, using the same procedure outlined above, we computed *weight gain ≥5%* in a 1-year period.

Longitudinal Weight Trajectory

Weight is frequently measured, often resulting in several weight measures per patient over time. Researchers may be interested in assessing weight trajectories within patients over time and potentially classifying patients according to their trajectory or examining whether types of patients respond differentially to interventions. Algorithm choice may affect the trajectory of individuals and their measurements collected over time, especially for algorithms that severely reduce the number of measurements left to analyze. Instead of aggregating patient weight over a specific period, studies analyzing weight measures within patients over time use repeated-measure designs such as (generalized) linear mixed models (LMMs), analysis of variance, or analysis of covariance for estimation. To compare algorithms in this context, we used a latent class LMM that assumes the population is heterogeneous and composed of some selected number of latent classes characterized by specific trajectories.

The latent class mixed models implemented through the R package *lcmm* (package version 1.8.1) [51] exhibited poor or slow convergence characteristics as the sample size increased; thus, a random sample of 1000 individuals from each cohort was used for model development. The same random sample was processed by each of the 12 algorithms and evaluated using the same latent class mixed model.

Facility-Level Metric

Researchers and evaluators are often interested in comparing facilities according to the percentage of patients who meet a metric of interest. To examine this application, we calculated the percentage of patients with 1-year *weight loss ≥5%* and *weight gain ≥5%* at each facility using the raw data and each of the 12 algorithms. Although these types of comparisons may often be risk-adjusted, our objective was only to understand the impact of algorithm choice on calculated facility-level metrics; therefore, we examined unadjusted facility rates. We rank-ordered facilities separately based on the percentage of patients with weight loss of ≥5% and weight gain of ≥5%. We then compared the differences in the percentage of patients based on each algorithm, grouping by those that used all data and period-specific algorithms.

Results

Sample Descriptive Statistics

Both cohorts included approximately 100,000 patients (n=98,786 in 2008 and n=99,958 in 2016; Multimedia Appendix 1, Table S2). Patients were excluded if they had no weight measurements or were pregnant during the data collection period.

Using the raw data from the 2016 cohort, each veteran had a mean of 12.2 (SD 24.9) weights recorded over the 4-year collection period, and 1 patient had 4981 measurements (web-based supplement [50]). Approximately 5.29% (5291/99,958) of veterans had only 1 weight measurement recorded. Before applying any cleaning rules, the data included 1,175,995 total weight measurements. Between 2008 and 2016, the average weight increased by approximately 2.3 kg (91.9-94.3 kg), with a 1-point increase in SD (21.6-22.0 kg; Multimedia Appendix 1, Table S3). The number of weights recorded did not differ between the 2008 and 2016 cohorts and had similar overall distributions.

Aside from the difference in average weight between the 2 cohorts, the results did not reveal major differences in the number of weight measurements per patient or weight distributions. Therefore, the remainder of the results will focus on the 2016 cohort. The results from the 2008 cohort are included in Multimedia Appendix 1.

Algorithm Descriptive Statistics

Descriptive statistics for the raw data and each of the 12 algorithms are shown in Table 2. After applying each algorithm to the raw data, all but 2 retained >90% of the patients—Kazerooni and Lim [23] retained approximately 24% (23,987/99,958) of patients, and Rosenberger et al [28] retained 63.43% (63,405/99,958). The mean and SD varied little between algorithm types, ranging from 93 to 95 kg (range 20.6-21.9 kg).

Table 2. Weight processing by algorithm and type of algorithm.

Item	Patients retained, n (% of raw weights)	Weight measurements retained, n (% of raw weights)	Weight (kg), mean (SD; range)	Weight (kg), median (IQR)
Raw weights	99,958 (100)	1,175,995 (100)	94.3 (22.0; 0-674.0)	91.8 (27.4)
Algorithms that used all data				
Buta et al [18]	90,159 (90.2)	1,131,996 (96.3)	94.3 (21.9; 12.3-111.1)	91.9 (27.3)
Chan and Raffa [19]	96,132 (96.2)	1,170,114 (99.5)	94.3 (21.9; 24.5-330.0)	91.8 (27.4)
Maguen et al [26]	98,352 (98.4)	1,037,293 (88.2)	93.3 (21.0; 31.9-245.4)	91.0 (26.4)
Breland et al [17]	99,958 (100)	1,175,177 (99.9)	94.3 (21.9; 34.0-315.0)	91.8 (27.4)
Maciejewski et al [25]	99,958 (100)	1,146,995 (97.5)	94.4 (21.8; 28.1-247.7)	91.9 (27.2)
Littman et al [24]	96,130 (96.2)	1,161,661 (98.8)	94.3 (21.8; 34.0-247.7)	91.9 (27.2)
Period-specific algorithms				
Rosenberger et al [28]	63,405 (63.4)	227,215 (19.3)	94.3 (21.0; 0-596.2)	92.0 (26.3)
Kazerooni and Lim [23]	23,987 (24)	71,961 (6.1)	94.8 (21.8; 0-559.6)	92.5 (27.2)
Goodrich et al [20]	95,748 (95.8)	199,830 (17)	93.5 (20.6; 36.3-226.8)	91.2 (25.7)
Janney et al [22]	95,742 (95.8)	199,830 (17)	93.5 (20.6; 35.6-247.7)	91.2 (25.7)
Jackson et al [21][a]	96,559 (96.6)	251,501 (21.4)	93.6 (20.6; 27.4-259.0)	91.2 (25.9)
Noël et al [27][a]	99,958 (100)	683,008 (58.1)	94.0 (20.9; 31.8-267.1)	91.6 (26.1)

[a]These algorithms differ from the other period-specific algorithms as they first use all available data and then proceed to aggregate measures by the mean or median within select periods.

The raw, unprocessed data contained implausible values ranging from 0 kg to 674 kg. Although most algorithms involved removing outlying values—often as the first step—some did not. Most notably, data processed by two of the algorithms (Kazerooni and Lim [23] and Rosenberger et al [28]) maintained weight values from 0 kg to >454 kg (see Table 1 for algorithm descriptions).

Algorithms designed to use all available weights retained a bulk of the measurements (1,037,293/1,175,995, 88.21% to 1,175,177/1,175,995, 99.93%) and resulted in a similar average weight (mean 93.3-94.4, SD 21.0-22.0 kg). The SD did not decrease after applying the algorithms except for the algorithm by Maguen et al [26], which retained 88.21% (1,037,293/1,175,995) of the measurements and resulted in a slightly lower average weight and SD (mean 93.3 kg, SD 21.0 kg).

For the period-specific algorithms, only 1 retained >50% of the raw weight measurements (Noël et al [27] maintained 683,008/1,175,995, 58.08% of the available data), yet the average weight and SDs differed little between algorithms. The Kazerooni and Lim [23] algorithm resulted in higher average and median weights (mean 94.8 kg, SD 21.8 kg, median 92.5 kg). It is important to note that the algorithms designed by

Kazerooni and Lim [23] and Noël et al [27] first use all available data and then proceed to aggregate measures by the mean or median within select periods. Thus, they differ in approach from the other period-specific algorithms, which first define periods and then extract weight measures during windows around those periods.

Although the mean weight did not change appreciably between the 12 algorithms, there were noticeable differences in the resulting distributions of weight. To explore these differences, we implemented a bootstrap procedure for the mean and variance by sampling 1000 patients, with replacement—thus each patient could be in each sampling iteration more than once—then evaluating the sample data with all 12 algorithms, and repeating this procedure 100 times. Each algorithm is designed to *clean* weight measurements; thus, in terms of the mean, the differences between algorithms are minute (Multimedia Appendix 1, Figure S1), rarely deviating from the mean of the unprocessed data. Differences in variance stand in stark contrast, deviating in both measures of center and spread between algorithms and years—most notably, Kazerooni and Lim [23] and Maguen et al [26] (Figure 1 [17-28]). Disregarding the standout algorithms, differences in SD were still small on an absolute scale, with an approximate range between algorithms of 0.9 kg and 1.8 kg.

Figure 1. Bootstrapped 95% CI of the SD by algorithm and algorithm type. The midpoint represents the median SD, the thick gray line represents the 80% quantile interval, and the black line represents the 95% quantile interval [23-25,30,35,36,38,39,41-43,46].
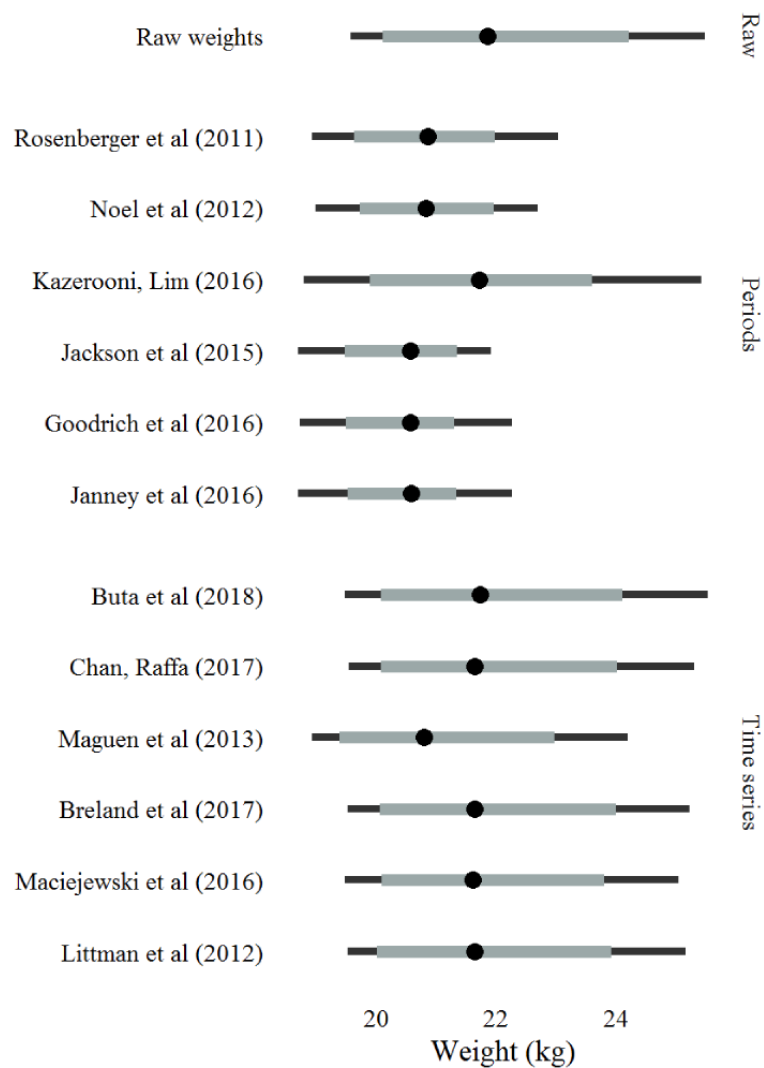
Algorithms Applied to Analysis Scenarios

Weight as a Predictor

A total of 13 individual logistic regressions were computed to predict the occurrence of new-onset diabetes as a function of weight. The reported OR and 95% CI varied little between algorithms, and all ORs were slightly >1.00 (Figure 2 [17-28]). The results from the Kazerooni and Lim [23] algorithm are the most striking, exhibiting the widest CI and the smallest OR (see the web-based supplement section *Weight as a Predictor* for a detailed exploration of each analytic decision [50]).

Figure 2. Odds ratio (OR; 95% CI) from 13 separate logistic regressions predicting new-onset diabetes as a function of weight [23-25,30,35,36,38,39,41-43,46].
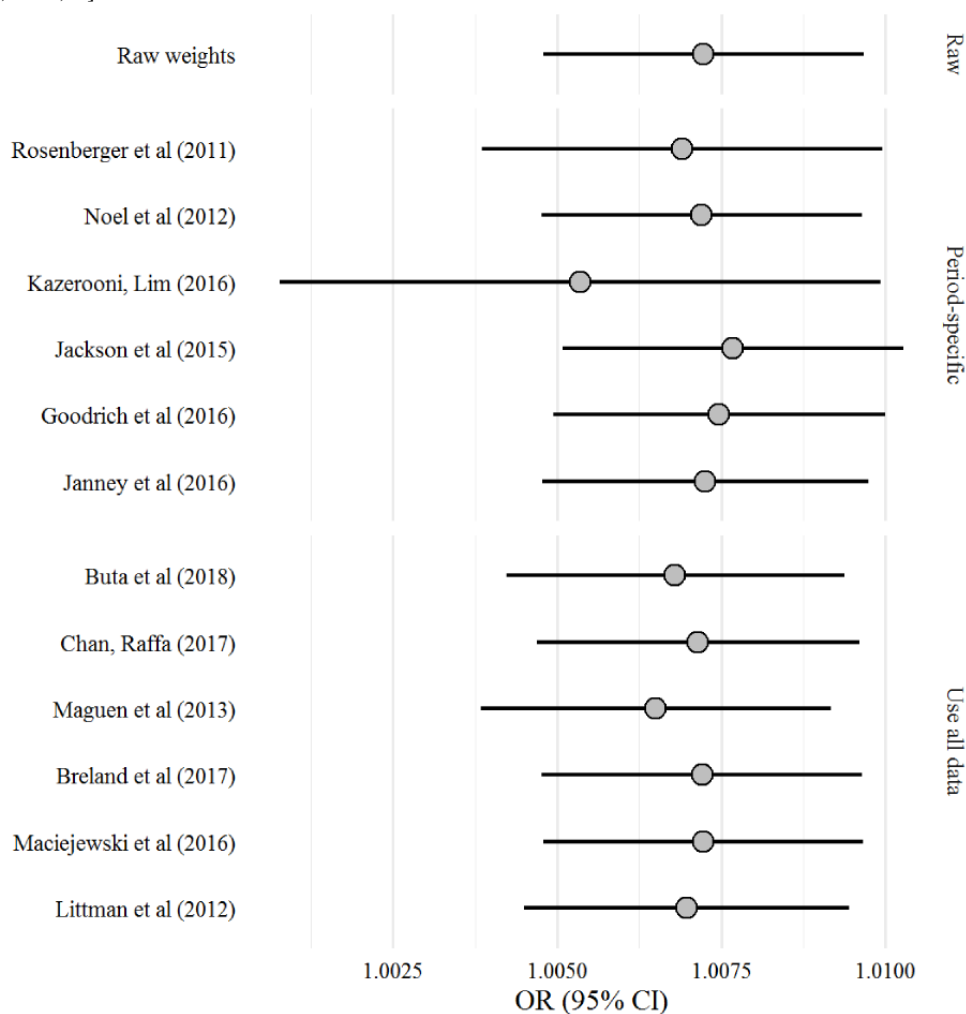
Weight Change (Gain and Loss)

Table 3 shows descriptive statistics for ≥5% weight loss and gain by algorithm. To calculate 1-year weight loss and gain, patients were required to have both a baseline weight (60 days before to 60 days after the index date) and a follow-up weight (60 days before to 60 days after 1 year from the index date). After applying the algorithms, only 24% (23,987/99,958) to 60.25% (60,225/99,958) of patients were retained for analysis and, unsurprisingly, the algorithms that used all data retained the most patients. However, the proportion of patients with ≥5% weight loss remained stationary at roughly 13.13% (7851/59,773) to 13.95% (5425/38,875) for nearly all algorithms. The exception was the Maguen et al [26] algorithm, which resulted in only 9.37% (4933/52,642) of patients

achieving this weight loss goal. A similar pattern in the results is exhibited by the weight gain analysis—Maguen et al [26] resulted in the lowest gain (4088/52,642, 7.77%), whereas all others stayed relatively the same at 10.86% (6494/59,770) to 12.15% (4725/38,875).

The average weight change was slightly <0, ranging from −0.13 kg to −0.43 kg, with the largest discrepancy resulting from the Maguen et al [26] algorithm and the smallest from the Kazerooni and Lim [23] algorithm. Despite the often lengthy processing steps involved in each algorithm, almost all algorithms still retained implausible weight change outliers, ranging from −1454 kg to −242 kg for the Rosenberger et al [28] and Kazerooni and Lim [23] algorithms, respectively (see Multimedia Appendix 1, Figure S2 for a graphical representation).

Table 3. Comparing weight loss metrics by algorithm, common measures of weight loss ≥5%, and average weight change from baseline.

Item	Patients retained[a], n (%)	Weight loss ≥5% from baseline, n (%)	Weight gain ≥5% from baseline, n (%)	Average weight change from baseline (kg), mean (SD; range)
Raw weights	60,286 (60.3)	8162 (13.5)	6977 (11.6)	–0.13 (7.3; –456 to –485)
Algorithms that used all data				
Buta et al [18]	57,014 (57)	7762 (13.6)	6642 (11.6)	–0.27 (5.4; –111 to –126)
Chan and Raffa [19]	60,175 (60.2)	8069 (13.4)	6902 (11.5)	–0.26 (5.4; –231 to –126)
Maguen et al [26]	52,642 (52.7)	4933 (9.4)	4088 (7.8)	–0.17 (3.5; –33 to –44)
Breland et al [17]	60,225 (60.3)	8124 (13.5)	6936 (11.5)	–0.27 (5.2; –117 to –94)
Maciejewski et al [25]	58,457 (58.5)	7985 (13.7)	6810 (11.6)	–0.28 (5.1; –53 to –88)
Littman et al [24]	59,773 (59.8)	7851 (13.1)	6787 (11.4)	–0.22 (4.9; –54 to –49)
Period-specific algorithms				
Rosenberger et al [28]	38,875 (38.9)	5425 (14)	4725 (12.2)	–0.31 (6.4; –454 to –135)
Kazerooni and Lim [23]	23,987 (24)	3355 (14)	2503 (10.4)	–0.43 (5.6; –242 to –136)
Goodrich et al [20]	58,142 (58.2)	7828 (13.5)	6688 (11.5)	–0.27 (5.2; –53 to –93)
Janney et al [22]	58,171 (58.2)	7842 (13.5)	6679 (11.5)	–0.28 (5.4; –132 to –127)
Jackson et al [21]	59,770 (59.8)	7973 (13.3)	6494 (10.9)	–0.32 (5.1; –111 to –104)
Noël et al [27]	58,525 (58.5)	7786 (13.3)	6624 (11.3)	–0.26 (5.2; –111 to –88)

[a]Number of patients retained after applying the algorithm. N=99,958 (number of veterans in the 2016 cohort).

Weight Trajectory

For each algorithm, the individual trajectories were modeled using a random slope and intercept. Latent class membership represents a choice by the statistical modeler; here, for both conceptual and parsimonious reasons, a 3-class model was chosen for analysis.

Figure 3 [17-28] displays predictions from latent class LMMs computed for each algorithm. There are three types of trajectories: those displayed with a negative slope (predicted weight loss), a slope of nearly 0 (corresponding to those predicted to maintain weight across a 1-year span), and a positive slope (predicted weight gain).

The choice of algorithm can affect predicted weight loss and weight gain within 1 year. Each algorithm produced a slightly different slope and intercept for each class (eg, for the raw data,

$\beta_{1,t}$=–0.00906 vs $\beta_{2,t}$=.00000 and $\beta_{3,t}$=.01322 for classes 1, 2, and 3, respectively), implying that the second class of individuals maintained their weight over time, whereas class 1 was predicted to lose 1.5 kg, and class 3 was predicted to gain 2.2 kg over a period of 365 days. For all but 1 algorithm, the posterior probability of individuals classified as class 1 (loss) was low, with a median across algorithms of 0.34 (range 0.014-0.99; Multimedia Appendix 1, Table S13), implying that 3.4% (34/1000) of sampled veterans were predicted to lose weight. The Kazerooni and Lim [23] algorithm differed and classified 99.6% (262/263) of its patients as class 1 (loss), with almost 0% predicted for class 2 (maintenance). Goodrich et al [20], Janney et al [22], Kazerooni and Lim [23], and Rosenberger et al [28] stand out with the steepest slopes in class 3 (gain), indicating greater predicted weight gain for patients in this class.

Figure 3. Group-based trajectory modeling by algorithm [23-25,30,35,36,38,39,41-43,46].
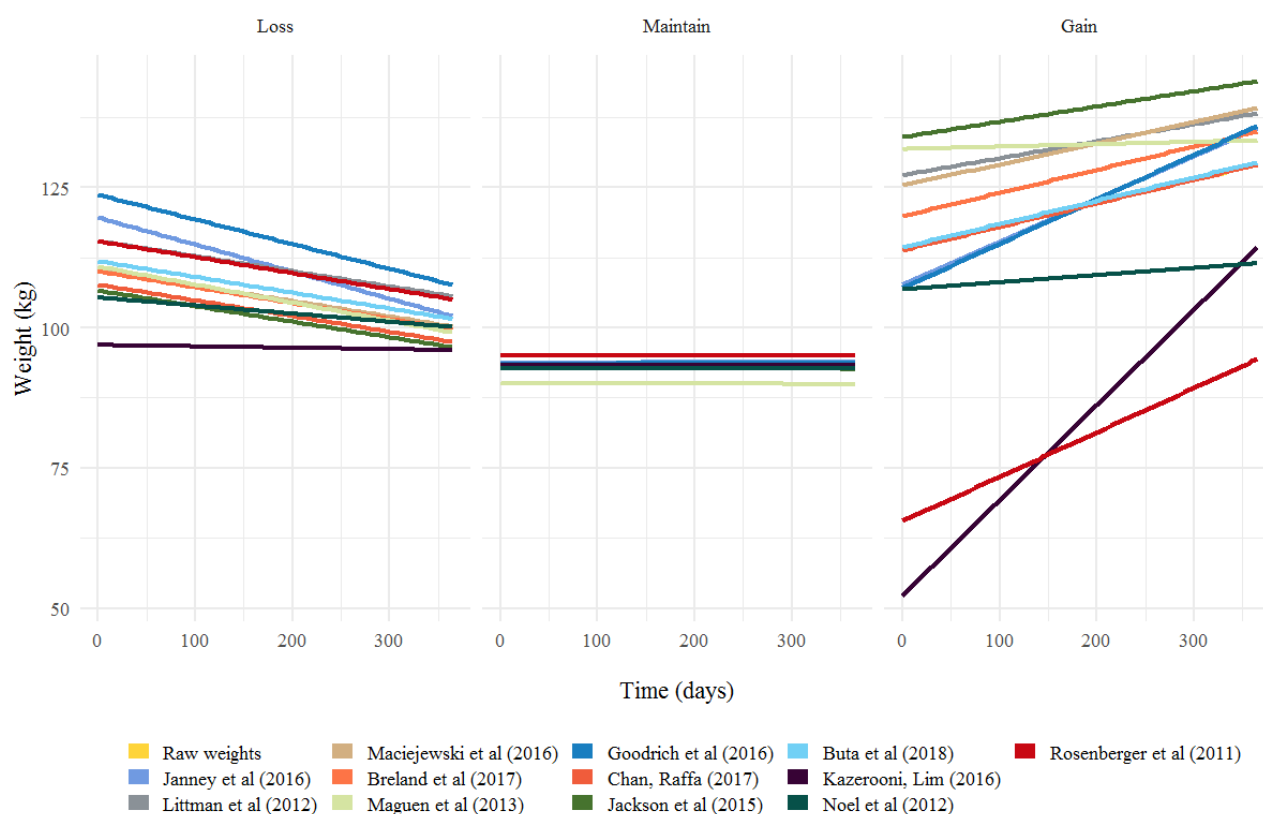
Facility-Level Metrics

The percentage of patients with ≥5% weight loss and gain was calculated for each of the 130 facilities using the raw weight data and the weight data as processed by each algorithm. Using the raw data, the percentage of patients with ≥5% weight loss ranged from 2% (1/44) to 19.7% (78/395) across facilities, with an average of 13.5% (SD 2.6%). Across algorithms, the percentage of patients who met the metric ranged from a minimum of 2% (1/44) to a maximum of 26% (13/50). For weight gain, the percentage of patients with ≥5% weight gain ranged from 6% (14/234) to 20% (9/44) across facilities using the raw data, with an average of 11.6% (SD 2.3%); across

algorithms, the percentage of patients ranged from 3.1% (12/386) to 27% (14/51). Figure 4 [17-28] shows the facility-level rates, with facilities ranked along the x-axis according to the percentage of patients who met the metric using raw data. Higher-ranking facilities had greater rates of patients meeting the metric. Using the period-specific algorithms to define the percentage of patients with ≥5% weight loss resulted in more variability, and the choice of algorithm clearly affected facility rank. In contrast, the algorithms that used all data exhibited similar ranking to the raw data. The Maguen et al [26] algorithm was a clear outlier and resulted in much lower rates that would affect facility ranking. The Maciejewski et al [25] algorithm showed slightly higher rates.

Figure 4. Facility-level percentage of patients with ≥5% weight loss by algorithm. Facilities are ranked along the x-axis according to the percentage of patients who met the metric using raw data, with higher-ranking facilities having greater rates of patients meeting the metric. The percentage of patients who met the metric calculated by each algorithm is displayed for each facility [23-25,30,35,36,38,39,41-43,46].
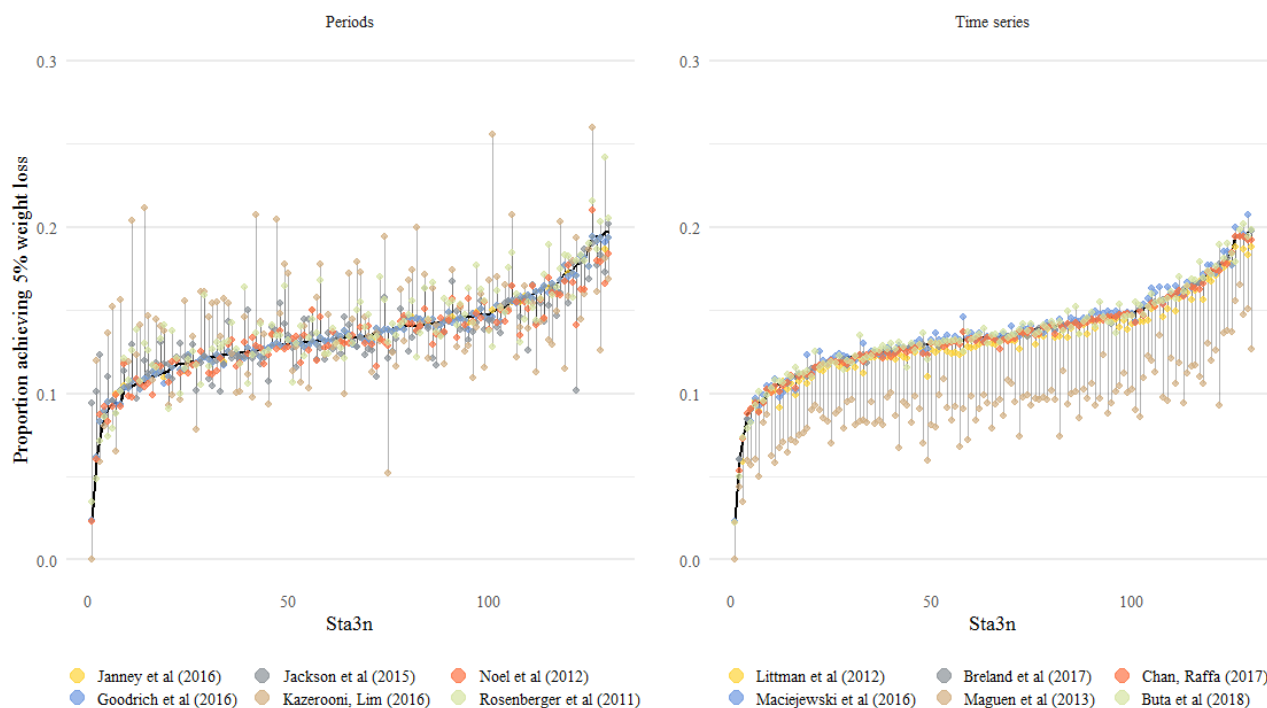
Discussion

Principal Findings

For many applications, the differences between weight-processing algorithms are minor, implying that a simpler algorithm design may be accurate and computationally more efficient in many scenarios. Furthermore, in some cases, the results are not appreciably different from using raw, unprocessed data.

There are subtleties between each algorithm and algorithm type that appear to be more appropriate for specific applications. For example, if it is assumed within a cohort that weight will be lost or gained linearly (eg, weight loss programs or patients with terminal cancer), the Maguen et al [26] algorithm would be appropriate to use.

Studies using point estimates of weight (descriptive statistics and weight as a predictor) and weight change may benefit from a simple cleaning rule based on cutoffs of implausible values, such as excluding weights <34 kg or >318 kg. However, we also recommend examining the computed *weight change* (output) for implausible values in addition to filtering the unprocessed measurements.

Among the algorithms that used all weight measures, most removed outliers within patients, often using some variation of *rolling* SDs to determine implausible values. However, the results from the study by Buta et al [18] are consistent with these algorithms even though the algorithms simply apply an outlier filter based on BMI to the entire sample.

Studies examining weight trajectories and facility-level metrics may benefit from a more nuanced algorithm that considers all available weight data. With respect to trajectory analyses, Kazerooni and Lim [23] and Janney et al [22], both period-specific algorithms, showed steeper weight losses and thus inconsistent results compared with other algorithms. Clearly, when modeling trajectories, the estimation would benefit from using an algorithm that uses all available weight data. In terms of facility-level analysis, all period-specific algorithms resulted in inconsistent or noisy results in comparison with the algorithms that used all data. The clear exception was the Maguen et al [26] algorithm, which *assumes* linearity in weight over time when cleaning weight measurements, an assumption that may not be tenable.

As an example of a recommendation, based on preliminary findings, we used a 2-stage algorithm to derive and clean a weight outcome for the study by Miech et al [52], specifically ≥5% weight loss in a 1-year time frame. The procedure used to arrive at the final outcome was as follows: for each patient in the VHA-derived cohort, all weight data were collected between a patient's *baseline* time point and the end of follow-up (1 year). To clean these data, the Breland et al [17] algorithm was used as it uses all data, shows consistent results in comparison with other algorithms that use all data, and provides a reasonable distribution of weight values upon computing weight change. Alternatively, the Maciejewski et al [25] algorithm could have been chosen as it exhibits the same ideal characteristics as the Breland et al [17] algorithm yet comes with added complexity in terms of parameter settings because of its design expectant of large changes in weight. Once cleaned with the Breland et al [17] algorithm, weight change and weight change as a percentage of body weight were calculated, and implausible values left in this distribution were then assessed iteratively by choosing the next closest measurement to either the baseline or

follow-up weight and then re-examining the weight change distribution. This process ended when the distribution was removed of all implausible values given a range chosen by the study investigators.

Considerations

These data can be stratified in many ways and, for the purposes of brevity, we chose to display the results assuming homogeneity of the sample. Alternatively, stratifying by demographic or clinical factors had the potential to change our results and conclusions; thus, we chose to differentiate our analysis for patient sex and for categories of weight—namely, underweight, overweight, and obese (web-based supplement [50]). For the sex subanalysis, the patterns of postalgorithm measurements did not differ between men and women save for the noisy facility-level analysis, which can be attributed to the small number of women in multiple facilities. A similar result can be seen in the analysis by BMI category, where the patterns were similar, but the facility-level analysis was noisy because of small numbers. Consequently, the value in further subanalyses should be explored to better address common clinical and research scenarios.

Similar to the choice of data, the methods we chose to address the impact of algorithms were tested on a small selection of analytic approaches while disregarding others that researchers may wish to use. Chiefly, we did not examine the impact on a broader set of machine learning or artificial or computational intelligence approaches common in big data analytics. Further combining machine learning, missing data imputation, and the impact of algorithm choice could prove to be an invaluable resource for the clinical research community.

Limitations

Our data lack a gold standard and thus, we cannot establish that a presumed outlier is in fact implausible; it is possible that some individuals experienced drastic weight changes that were not considered. Patients who were pregnant during the period were excluded; however, other diseases or conditions may be associated with dramatic weight shifts, and amputation in diabetic patients could also be considered. We did examine the impact of including weight measures from the inpatient setting as well as bariatric surgery patients and found only 2 individuals

with implausible weight change values (web-based supplement [50]).

In addition, many algorithms were designed using a specific cohort of patients or an analytic approach, which may not transfer to a general patient cohort. The Maciejewski et al [25] algorithm was designed specifically for studies involving patients who had undergone bariatric surgery or patients who experienced drastic weight changes within a short amount of time. Furthermore, Noël et al [27] proceeded by aggregating longitudinally measured weight over fiscal quarters, a method more appropriate for econometric-type research studies.

Our conclusion that applying a simple algorithm or filter may be enough to *clean* the data has been arrived at by analyzing large samples; thus, these results may differ in smaller samples or small subpopulations, as can be seen in the sex and BMI category analyses. We did not analyze the differences in the algorithms because of the sample size in this study. A simulation study would be warranted to fully assess the impact of sample size.

Finally, all algorithms were reconstructed from the published methods and supplemental material, and there was potential for misinterpretation. In the era of big data analytics and use of patient EHR data for research and evaluation, it is essential that details surrounding data processing and measure creation are included in supplemental materials or shared code (eg, GitHub, Bitbucket, or Docker) to facilitate reproducibility and replication efforts.

Conclusions

In this paper, we presented several applications of algorithms to process weight measurements obtained from EHRs and attempted to provide recommendations for common research scenarios. Different algorithms result in generally similar results. In some cases, the results are not different from using raw, unprocessed data, despite algorithm complexity. Studies using point estimates of weight (descriptive statistics and weight as a predictor) and weight change may benefit from a simple cleaning rule based on cutoffs of implausible values. Research questions involving weight trajectories and facility-level metrics may benefit from a more nuanced algorithm that considers all available weight data.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Supplementary tables and figures.
[DOCX File , 376 KB - medinform_v10i3e30328_app1.docx]

References

1. Mennemeyer ST, Menachemi N, Rahurkar S, Ford EW. Impact of the HITECH Act on physicians' adoption of electronic health records. J Am Med Inform Assoc 2016 Mar;23(2):375-379. [doi: 10.1093/jamia/ocv103] [Medline: 26228764]

2. Hripcsak G, Albers DJ. Next-generation phenotyping of electronic health records. J Am Med Inform Assoc 2013 Jan 01;20(1):117-121 [FREE Full text] [doi: 10.1136/amiajnl-2012-001145] [Medline: 22955496]

3. Tate AR, Beloff N, Al-Radwan B, Wickson J, Puri S, Williams T, et al. Exploiting the potential of large databases of electronic health records for research using rapid search algorithms and an intuitive query interface. J Am Med Inform Assoc 2014;21(2):292-298 [FREE Full text] [doi: 10.1136/amiajnl-2013-001847] [Medline: 24272162]

4. Zozus M, Richesson R, Hammond W. Acquiring and using electronic health record data. NIH Collaboratory Electronic Health Records. URL: https://rethinkingclinicaltrials.org/resources/acquiring-and-using-electronic-health-record-data/ [accessed 2022-02-12]

5. Goldstein BA, Navar AM, Pencina MJ, Ioannidis JP. Opportunities and challenges in developing risk prediction models with electronic health records data: a systematic review. J Am Med Inform Assoc 2017 Jan;24(1):198-208 [FREE Full text] [doi: 10.1093/jamia/ocw042] [Medline: 27189013]

6. Springate DA, Kontopantelis E, Ashcroft DM, Olier I, Parisi R, Chamapiwa E, et al. ClinicalCodes: an online clinical codes repository to improve the validity and reproducibility of research using electronic medical records. PLoS One 2014;9(6):e99825 [FREE Full text] [doi: 10.1371/journal.pone.0099825] [Medline: 24941260]

7. Weiskopf NG, Weng C. Methods and dimensions of electronic health record data quality assessment: enabling reuse for clinical research. J Am Med Inform Assoc 2013 Jan 01;20(1):144-151 [FREE Full text] [doi: 10.1136/amiajnl-2011-000681] [Medline: 22733976]

8. Williams R, Kontopantelis E, Buchan I, Peek N. Clinical code set engineering for reusing EHR data for research: a review. J Biomed Inform 2017 Jun;70:1-13 [FREE Full text] [doi: 10.1016/j.jbi.2017.04.010] [Medline: 28442434]

9. Gulliford MC, Charlton J, Ashworth M, Rudd AG, Toschke AM, eCRT Research Team. Selection of medical diagnostic codes for analysis of electronic patient records. Application to stroke in a primary care database. PLoS One 2009 Sep 24;4(9):e7168 [FREE Full text] [doi: 10.1371/journal.pone.0007168] [Medline: 19777060]

10. Jensen MD, Ryan DH, Apovian CM, Ard JD, Comuzzie AG, Donato KA, American College of Cardiology/American Heart Association Task Force on Practice Guidelines, Obesity Society. 2013 AHA/ACC/TOS guideline for the management of overweight and obesity in adults: a report of the American College of Cardiology/American Heart Association Task Force on Practice Guidelines and The Obesity Society. Circulation 2014 Jun 24;129(25 Suppl 2):S102-S138 [FREE Full text] [doi: 10.1161/01.cir.0000437739.71477.ee] [Medline: 24222017]

11. Diabetes Prevention Program Research Group. Long-term effects of lifestyle intervention or metformin on diabetes development and microvascular complications over 15-year follow-up: the Diabetes Prevention Program Outcomes Study. Lancet Diabetes Endocrinol 2015 Nov;3(11):866-875 [FREE Full text] [doi: 10.1016/S2213-8587(15)00291-0] [Medline: 26377054]

12. Finkelstein EA, Trogdon JG, Cohen JW, Dietz W. Annual medical spending attributable to obesity: payer-and service-specific estimates. Health Aff (Millwood) 2009;28(5):w822-w831. [doi: 10.1377/hlthaff.28.5.w822] [Medline: 19635784]

13. Annis A, Freitag MB, Evans RR, Wiitala WL, Burns J, Raffa SD, et al. Construction and use of body weight measures from administrative data in a large national health system: a systematic review. Obesity (Silver Spring) 2020 Jul;28(7):1205-1214 [FREE Full text] [doi: 10.1002/oby.22790] [Medline: 32478469]

14. VIReC Research User Guides. U.S. Department of Veterans Affairs. URL: https://www.virec.research.va.gov/Resources/RUGs.asp [accessed 2022-02-12]

15. Wiitala W, Vincent B, Burns J, Prescott HC, Waljee AK, Cohen GR, et al. Variation in laboratory test naming conventions in EHRs within and between hospitals: a nationwide longitudinal study. Med Care 2019 Apr;57(4):e22-e27 [FREE Full text] [doi: 10.1097/MLR.0000000000000996] [Medline: 30394981]

16. VHA OIA data quality program. VHA Data Quality Program Laboratory Reports. URL: http://vaww.vhadataportal.med.va.gov/Resources/DataReports.aspx [accessed 2022-02-25]

17. Breland JY, Phibbs CS, Hoggatt KJ, Washington DL, Lee J, Haskell S, et al. The obesity epidemic in the veterans health administration: prevalence among key populations of women and men veterans. J Gen Intern Med 2017 Apr;32(Suppl 1):11-17 [FREE Full text] [doi: 10.1007/s11606-016-3962-1] [Medline: 28271422]

18. Buta E, Masheb R, Gueorguieva R, Bathulapalli H, Brandt CA, Goulet JL. Posttraumatic stress disorder diagnosis and gender are associated with accelerated weight gain trajectories in veterans during the post-deployment period. Eat Behav 2018 Apr;29:8-13 [FREE Full text] [doi: 10.1016/j.eatbeh.2018.01.002] [Medline: 29413821]

19. Chan SH, Raffa SD. Examining the dose-response relationship in the veterans health administration's MOVE! Weight management program: a nationwide observational study. J Gen Intern Med 2017 Apr;32(Suppl 1):18-23 [FREE Full text] [doi: 10.1007/s11606-017-3992-3] [Medline: 28271425]

20. Goodrich DE, Klingaman EA, Verchinina L, Goldberg RW, Littman AJ, Janney CA, et al. Sex differences in weight loss among veterans with serious mental illness: observational study of a national weight management program. Womens Health Issues 2016;26(4):410-419. [doi: 10.1016/j.whi.2016.05.001] [Medline: 27365284]

21. Jackson SL, Long Q, Rhee MK, Olson DE, Tomolo AM, Cunningham SA, et al. Weight loss and incidence of diabetes with the Veterans Health Administration MOVE! lifestyle change programme: an observational study. Lancet Diabetes Endocrinol 2015 Mar;3(3):173-180. [doi: 10.1016/S2213-8587(14)70267-0]

22. Janney CA, Kilbourne AM, Germain A, Lai Z, Hoerster KD, Goodrich DE, et al. The influence of sleep disordered breathing on weight loss in a national weight management program. Sleep 2016 Jan 01;39(1):59-65 [FREE Full text] [doi: 10.5665/sleep.5318] [Medline: 26350475]

23. Kazerooni R, Lim J. Topiramate-associated weight loss in a veteran population. Military Med 2016 Mar;181(3):283-286. [doi: 10.7205/milmed-d-14-00636]

24. Littman AJ, Boyko EJ, McDonell MB, Fihn SD. Evaluation of a weight management program for veterans. Prev Chronic Dis 2012;9:E99 [FREE Full text] [doi: 10.5888/pcd9.110267] [Medline: 22595323]

25. Maciejewski ML, Arterburn DE, Van Scoyoc L, Smith VA, Yancy WS, Weidenbacher HJ, et al. Bariatric surgery and long-term durability of weight loss. JAMA Surg 2016 Nov 01;151(11):1046-1055 [FREE Full text] [doi: 10.1001/jamasurg.2016.2317] [Medline: 27579793]

26. Maguen S, Madden E, Cohen B, Bertenthal D, Neylan T, Talbot L, et al. The relationship between body mass index and mental health among Iraq and Afghanistan veterans. J Gen Intern Med 2013 Jul;28 Suppl 2:S563-S570 [FREE Full text] [doi: 10.1007/s11606-013-2374-8] [Medline: 23807066]

27. Noël PH, Wang CP, Bollinger MJ, Pugh MJ, Copeland LA, Tsevat J, et al. Intensity and duration of obesity-related counseling: association with 5-Year BMI trends among obese primary care patients. Obesity (Silver Spring) 2012 Apr;20(4):773-782 [FREE Full text] [doi: 10.1038/oby.2011.335] [Medline: 22134198]

28. Rosenberger PH, Ning Y, Brandt C, Allore H, Haskell S. BMI trajectory groups in veterans of the Iraq and Afghanistan wars. Prev Med 2011 Sep;53(3):149-154 [FREE Full text] [doi: 10.1016/j.ypmed.2011.07.001] [Medline: 21771610]

29. Adams CE, Gabriele JM, Baillie LE, Dubbert PM. Tobacco use and substance use disorders as predictors of postoperative weight loss 2 years after bariatric surgery. J Behav Health Serv Res 2012 Oct;39(4):462-471. [doi: 10.1007/s11414-012-9277-z] [Medline: 22374227]

30. Arterburn D, Livingston EH, Olsen MK, Smith VA, Kavee AL, Kahwati LC, et al. Predictors of initial weight loss after gastric bypass surgery in twelve Veterans Affairs Medical Centers. Obes Res Clin Pract 2013;7(5):e367-e376. [doi: 10.1016/j.orcp.2012.02.009] [Medline: 24304479]

31. Baker JF, Cannon GW, Ibrahim S, Haroldsen C, Caplan L, Mikuls TR. Predictors of longterm changes in body mass index in rheumatoid arthritis. J Rheumatol 2015 Jun;42(6):920-927 [FREE Full text] [doi: 10.3899/jrheum.141363] [Medline: 25834210]

32. Batch BC, Goldstein K, Yancy WS, Sanders LL, Danus S, Grambow SC, et al. Outcome by gender in the veterans health administration motivating overweight/obese veterans everywhere weight management program. J Womens Health (Larchmt) 2018 Jan;27(1):32-39 [FREE Full text] [doi: 10.1089/jwh.2016.6212] [Medline: 28731844]

33. Bounthavong M, Tran J, Golshan S, Piland N, Morello C, Blickensderfer A, et al. Retrospective cohort study evaluating exenatide twice daily and long-acting insulin analogs in a Veterans Health Administration population with type 2 diabetes. Diabetes Metab 2014 Sep;40(4):284-291. [doi: 10.1016/j.diabet.2014.06.002] [Medline: 25059703]

34. Braun K, Erickson M, Utech A, List R, Garcia JM. Evaluation of Veterans MOVE! Program for weight loss. J Nutr Educ Behav 2016 May;48(5):299-303.e1. [doi: 10.1016/j.jneb.2016.02.012] [Medline: 27169639]

35. Copeland LA, Pugh MJ, Hicks PB, Noel PH. Use of obesity-related care by psychiatric patients. Psychiatr Serv 2012 Mar;63(3):230-236. [doi: 10.1176/appi.ps.201100221] [Medline: 22307880]

36. Garvin JT. Weight reduction goal achievement with high-intensity MOVE!(®) treatment. Public Health Nurs 2015;32(3):232-236. [doi: 10.1111/phn.12194] [Medline: 25809093]

37. Garvin JT, Marion LN, Narsavage GL, Finnegan L. Characteristics influencing weight reduction among veterans in the MOVE!® Program. West J Nurs Res 2015 Jan;37(1):50-65. [doi: 10.1177/0193945914534323] [Medline: 24842681]

38. Garvin JT, Hardy D, Xu H. Initial response to program, program participation, and weight reduction among 375 move! Participants, Augusta, Georgia, 2008-2010. Prev Chronic Dis 2016 Apr 21;13:E55 [FREE Full text] [doi: 10.5888/pcd13.150598] [Medline: 27103265]

39. Grabarczyk TR. Observational comparative effectiveness of pharmaceutical treatments for obesity within the veterans health administration. Pharmacotherapy 2018 Jan;38(1):19-28. [doi: 10.1002/phar.2048] [Medline: 29044720]

40. Hoerster KD, Lai Z, Goodrich DE, Damschroder LJ, Littman AJ, Klingaman EA, et al. Weight loss after participation in a national VA weight management program among veterans with or without PTSD. Psychiatr Serv 2014 Nov 01;65(11):1385-1388. [doi: 10.1176/appi.ps.201300404] [Medline: 25123784]

41. Huizinga MM, Roumie CL, Greevy RA, Liu X, Murff HJ, Hung AM, et al. Glycemic and weight changes after persistent use of incident oral diabetes therapy: a Veterans Administration retrospective cohort study. Pharmacoepidemiol Drug Saf 2010 Nov;19(11):1108-1112. [doi: 10.1002/pds.2035] [Medline: 20878643]

42. Ikossi DG, Maldonado JR, Hernandez-Boussard T, Eisenberg D. Post-traumatic stress disorder (PTSD) is not a contraindication to gastric bypass in veterans with morbid obesity. Surg Endosc 2010 Aug;24(8):1892-1897. [doi: 10.1007/s00464-009-0866-8] [Medline: 20063014]

43. Kahwati LC, Lewis MA, Kane H, Williams PA, Nerz P, Jones KR, et al. Best practices in the Veterans Health Administration's MOVE! Weight management program. Am J Prev Med 2011 Nov;41(5):457-464. [doi: 10.1016/j.amepre.2011.06.047] [Medline: 22011415]

44. Littman AJ, Damschroder LJ, Verchinina L, Lai Z, Kim HM, Hoerster KD, et al. National evaluation of obesity screening and treatment among veterans with and without mental health disorders. Gen Hosp Psychiatry 2015;37(1):7-13. [doi: 10.1016/j.genhosppsych.2014.11.005] [Medline: 25500194]

45. Pandey N, Ashfaq SN, Dauterive EW, MacCarthy AA, Copeland LA. Military sexual trauma and obesity among women veterans. J Womens Health (Larchmt) 2018 Mar;27(3):305-310. [doi: 10.1089/jwh.2016.6105] [Medline: 28880738]

46. Romanova M, Liang L, Deng ML, Li Z, Heber D. Effectiveness of the MOVE! Multidisciplinary weight loss program for veterans in Los Angeles. Prev Chronic Dis 2013 Jul 03;10:E112 [FREE Full text] [doi: 10.5888/pcd10.120325] [Medline: 23823701]

47. Rutledge T, Braden AL, Woods G, Herbst KL, Groesz LM, Savu M. Five-year changes in psychiatric treatment status and weight-related comorbidities following bariatric surgery in a veteran population. Obes Surg 2012 Nov;22(11):1734-1741. [doi: 10.1007/s11695-012-0722-0] [Medline: 23011461]

48. Shi L, Zhao Y, Szymanski K, Yau L, Fonseca V. Impact of thiazolidinedione safety warnings on medication use patterns and glycemic control among veterans with diabetes mellitus. J Diabetes Complications 2011;25(3):143-150. [doi: 10.1016/j.jdiacomp.2010.06.003] [Medline: 20708416]

49. Xiao DY, Luo S, O'Brian K, Liu W, Carson KR. Weight change trends and overall survival in United States veterans with follicular lymphoma treated with chemotherapy. Leuk Lymphoma 2017 Apr;58(4):851-858 [FREE Full text] [doi: 10.1080/10428194.2016.1217526] [Medline: 27669828]

50. DCEP Project - weight algorithm development, evaluation, and analyses. GitHub. URL: https://github.com/ccmrcodes/weightalgorithms [accessed 2022-02-12]

51. Proust-Lima C, Philipps V, Liquet B. Estimation of extended mixed models using latent classes and latent processes: the R package lcmm. J Stat Softw 2017;78(2):1-56. [doi: 10.18637/jss.v078.i02]

52. Miech EJ, Freitag MB, Evans RR, Burns JA, Wiitala WL, Annis A, et al. Facility-level conditions leading to higher reach: a configurational analysis of national VA weight management programming. BMC Health Serv Res 2021 Aug 11;21(1):797 [FREE Full text] [doi: 10.1186/s12913-021-06774-w] [Medline: 34380495]

Abbreviations

CDW: Corporate Data Warehouse
EHR: electronic health record
LMM: linear mixed model
OR: odds ratio
VHA: Veterans Health Administration

A Digital Screening System for Alzheimer Disease Based on a Neuropsychological Test and a Convolutional Neural Network: System Development and Validation

Wen-Ting Cheah[1], MSc; Jwu-Jia Hwang[1], MSc; Sheng-Yi Hong[1], MSc; Li-Chen Fu[1], PhD; Yu-Ling Chang[2], PhD; Ta-Fu Chen[3], MD; I-An Chen[4], BSc; Chun-Chen Chou[4], BSc

[1]Department of Computer Science and Information Engineering, National Taiwan University, Taipei, Taiwan

[2]Department of Psychology, National Taiwan University, Taipei, Taiwan

[3]Department of Neurology, National Taiwan University Hospital, Taipei, Taiwan

[4]Taipei City Zhishan Senior Home, Taipei, Taiwan

Corresponding Author:
Li-Chen Fu, PhD
Department of Computer Science and Information Engineering
National Taiwan University
No. 1, Sec. 4, Roosevelt Rd
Taipei, 10617
Taiwan
Phone: 886 935545846
Email: lichen@ntu.edu.tw

Abstract

Background: Alzheimer disease (AD) and other types of dementia are now considered one of the world's most pressing health problems for aging people worldwide. It was the seventh-leading cause of death, globally, in 2019. With a growing number of patients with dementia and increasing costs for treatment and care, early detection of the disease at the stage of mild cognitive impairment (MCI) will prevent the rapid progression of dementia. In addition to reducing the physical and psychological stress of patients' caregivers in the long term, it will also improve the everyday quality of life of patients.

Objective: The aim of this study was to design a digital screening system to discriminate between patients with MCI and AD and healthy controls (HCs), based on the Rey-Osterrieth Complex Figure (ROCF) neuropsychological test.

Methods: The study took place at National Taiwan University between 2018 and 2019. In order to develop the system, pretraining was performed using, and features were extracted from, an open sketch data set using a data-driven deep learning approach through a convolutional neural network. Later, the learned features were transferred to our collected data set to further train the classifier. The first data set was collected using pen and paper for the traditional method. The second data set used a tablet and smart pen for data collection. The system's performance was then evaluated using the data sets.

Results: The performance of the designed system when using the data set that was collected using the traditional pen and paper method resulted in a mean area under the receiver operating characteristic curve (AUROC) of 0.913 (SD 0.004) when distinguishing between patients with MCI and HCs. On the other hand, when discriminating between patients with AD and HCs, the mean AUROC was 0.950 (SD 0.003) when using the data set that was collected using the digitalized method.

Conclusions: The automatic ROCF test scoring system that we designed showed satisfying results for differentiating between patients with AD and MCI and HCs. Comparatively, our proposed network architecture provided better performance than our previous work, which did not include data augmentation and dropout techniques. In addition, it also performed better than other existing network architectures, such as AlexNet and Sketch-a-Net, with transfer learning techniques. The proposed system can be incorporated with other tests to assist clinicians in the early diagnosis of AD and to reduce the physical and mental burden on patients' family and friends.

XSL·FO
RenderX

Introduction

Background

According to the latest report from Alzheimer's Disease International [1], the number of people with dementia worldwide will increase from 50 million in 2019 to 152 million by 2050, and the global annual cost of dementia is estimated to increase from US $1 trillion in 2019 to US $2 trillion in 2030. Dementia is also the seventh-leading cause of death in the world [2]. These numbers continue to grow year by year, and the risk of developing dementia grows significantly with increasing age. Therefore, as more and more countries' aging populations increase, there is an urgent need to put more effort into research related to this issue, since there is no cure for AD and the existing treatment is to extend the period of rapid progression of the disease.

AD is the most common etiology associated with dementia, and it accounts for approximately 60% to 70% of all dementia cases [3]. AD caused by the destruction and death of neurons in the brain is a syndrome related to ongoing decline in cognitive function in domains such as memory, visuospatial processing, language, and executive function; this decline results in impairment in carrying out the instrumental and basic activities of daily living [4].

MCI is a transitional state between normal aging and dementia, in which a patient's cognitive function undergoes mild but perceptible decline, as shown in Figure 1 [5]. Such degradation of cognitive function occurs more quickly than in normal aging, but unlike in AD, it does not affect the patient's ability to handle daily activities. According to the updated American Academy of Neurology guideline on MCI [6], about 14.9% of patients with MCI older than 65 years of age developed dementia at a 2-year follow-up. In clinical trials involving patients with MCI who had memory loss, most of them who progressed to having dementia had AD.

Figure 1. The continuum of Alzheimer disease [5]. MCI: mild cognitive impairment.
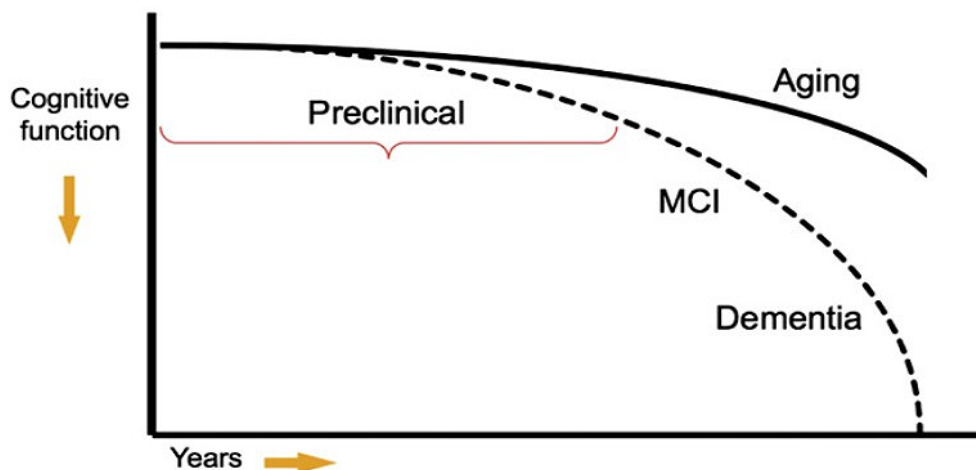
Currently, the diagnoses of the MCI and AD are based on the clinical judgment of doctors according to the symptoms, medical reports, and medical history from the individual, family members, friends, or caregivers. Additionally, a series of cognitive tests and neuropsychological assessments, such as the Mini-Mental State Examination (MMSE) [7] and the Clinical Dementia Rating scale [8], are essential to evaluate the individual's cognitive function. Furthermore, biomarker measurements that include cerebrospinal fluid testing and neuroimaging, such as structural magnetic resonance imaging (MRI) and positron emission tomography (PET), are also used to aid in diagnosis [9].

Several challenges need to be addressed to propose a screening system for the early detection of AD. One of the challenges is that the characteristics or signs of the early stage of the disease may not be obvious [10]. Moreover, the high cost of manual feature extraction needs to be avoided. However, meaningful feature representation has to be determined for building a screening model for the disease. As a screening tool, the efficiency of the overall screening process is another issue that needs to be considered. In this work, we proposed a digital screening system to reduce the burden on clinicians.

Purpose

The aim of this research was to propose a data-driven convolutional neural network (CNN) architecture through transfer learning and deep learning methods to discriminate between patients with AD or MCI and healthy controls (HCs). The designed CNN architecture was developed for a Rey-Osterrieth Complex Figure (ROCF) test system that automatically calculates scores to assist diagnosis. The purpose of the proposed system is to prevent late diagnosis of AD among older adults. Nevertheless, the proposed system will also reduce the manual workload for clinicians and diagnostic costs.

Related Work

Overview

When AD and other types of dementia collectively became one of the primary public health concerns worldwide, many different types of research studies began to develop diagnostic tools to

accurately classify individuals as having AD or MCI or as cognitively unimpaired individuals, also known as HCs. These studies can be categorized into two main types: neuroimaging studies and neuropsychological test studies.

Neuroimaging Studies

AD is a neurodegenerative disease, and the most remarkable brain changes appear to occur in the hippocampal formation and the entorhinal cortex, which are critical brain structures related to memory function. MRI is commonly used to measure the structural atrophy of the hippocampus and entorhinal cortex. Compared with cognitively unimpaired older adults and individuals with MCI, patients with AD have a smaller-sized hippocampus and entorhinal cortex [11]. Functional MRI provides information on the flow of oxygenated blood in the brain to detect higher brain cell activity by higher blood flow; it can be used to record the activation patterns of neural networks in the hippocampus when the participant is performing memory tests [12]. Furthermore, with the injection of a radioactive contrast agent into the human body, a PET scan can be used to obtain information on glucose metabolism and the brain's neurotransmitter activity.

With the help of multiscale feature extraction from baseline local hippocampus MRI data, Hu et al [13] adopted support vector machine (SVM) learning models to distinguish between patients with MCI that converted to AD and patients with MCI that did not convert to AD, and to distinguish between patients with AD and HCs. Challis et al [14] applied functional MRI scans and Bayesian Gaussian process logistic regression models to distinguish between HCs and patients with amnestic MCI, and between patients with amnestic MCI and those with AD. Li et al [15] used fusion information from MRI and PET scans for feature selection, processed the selected features through restricted Boltzmann machines to obtain the learned features, and applied the learned features to an SVM model for the classification of the different stages of AD. However, neuroimaging is not cheap. Moreover, patients who experience claustrophobia cannot undergo scanning by the machine because patients need to lie motionless inside the closed shell of the machine. Furthermore, patients with metallic implants, such as pacemakers, cannot undergo MRI due to the magnetic and radiofrequency fields generated during imaging. In addition, patients will be exposed to radiation while undergoing a PET scan.

Neuropsychological Test Studies

Neuropsychological assessments employing specifically designed tests are important for evaluating the brain dysfunction's behavioral and functional expression [16]. A neuropsychological test is typically administered to a participant by an examiner or neuropsychologist in a quiet environment. The purpose of the assessment is to gather the participant's cognitive and behavioral performance information. The MMSE is a widely used screening test for evaluating the cognitive status of older adults. However, it has limited utility in distinguishing between the patients with MCI and people in a standard aging group [17].

Drawing tests are widely used to assess constructional abilities, where the patient is asked to copy a complex figure and then recall and replicate the figure from memory. The Clock-Drawing Test (CDT) is a simple tool for screening people with dementia. It requires participants to draw the clock correctly using appropriate abilities, such as understanding language, planning, visualizing orientations, and executing the appropriate movement. However, people with dementia will not draw correctly due to impaired cognitive abilities, such as visual constructional processing, memory function, semantic knowledge retrieval, or executive function. Prange and Sonntag [18] proposed a digital CDT by implementing the Mendez scoring system [19] and creating a hierarchy of error categories to model the characteristics of CDT. Nevertheless, according to a survey report [20], many researchers using the CDT cannot significantly distinguish between patients with MCI and cognitively unimpaired participants, and the sensitivity and specificity have also been less satisfactory in most studies.

The ROCF test is widely used to assess visuospatial constructional capabilities and visual memory function [21]. It is a score-based neuropsychological assessment tool that assesses the individual's visual memory by testing their ability to draw a complex figure by copying, immediate recall, and delayed recall from memory. The ROCF test was first constructed by Rey [22], and it was then standardized by Osterrieth [23]. Miller et al [24] showed that combining the ROCF test with the MMSE can enhance the performance of the detection of individuals with MCI. Salvadori et al [25] evaluated the ROCF test using the Boston Qualitative Scoring System (BQSS) [26] in order to distinguish vascular MCI from degenerative MCI. There are several different scoring systems for quantifying the performance of the ROCF test, for example, the Rey-Osterrieth 36-point scoring system [27], the Developmental Scoring System [28], and the BQSS.

Nevertheless, the current scoring system of the ROCF test is labor intensive and needs to be performed by trained experts, due to the complexity of the scoring criterion. However, cognitive impairment in individuals with MCI is often subtle. It appears to be more challenging to distinguish between patients with MCI and HCs than between patients with AD and HCs, as the current manual scoring system may also have a limited ability to detect subtle differences between individuals with MCI and HCs.

Methods

System Overview

The proposed approach was partitioned conceptually into two portions, namely model training and screening, as depicted in Figure 2. The model training portion mainly included three modules: data collection, pretraining, and retraining. First, according to the standardized assessment protocol, participants had to be classified as individuals with MCI or HCs by an experienced doctor and neuropsychologist. Therefore, the diagnosis results were used to train the classification model as the ground truth. Second, we collected the ROCF test drawings from all participants, and a large, open, sketch data set was used for pretraining our proposed screening system. Third, the

XSL•FO

RenderX

screening model was implemented by applying the pretrained model to the collected data. Finally, we used the retraining model to discriminate participants. The screening portion used the system to classify new participants by differentiating cognitively unimpaired individuals from patients with AD or MCI. The following sections discuss the detailed implementation of each part in more detail.

Figure 2. System overview. AD: Alzheimer disease; HC: healthy control; MCI: mild cognitive impairment.
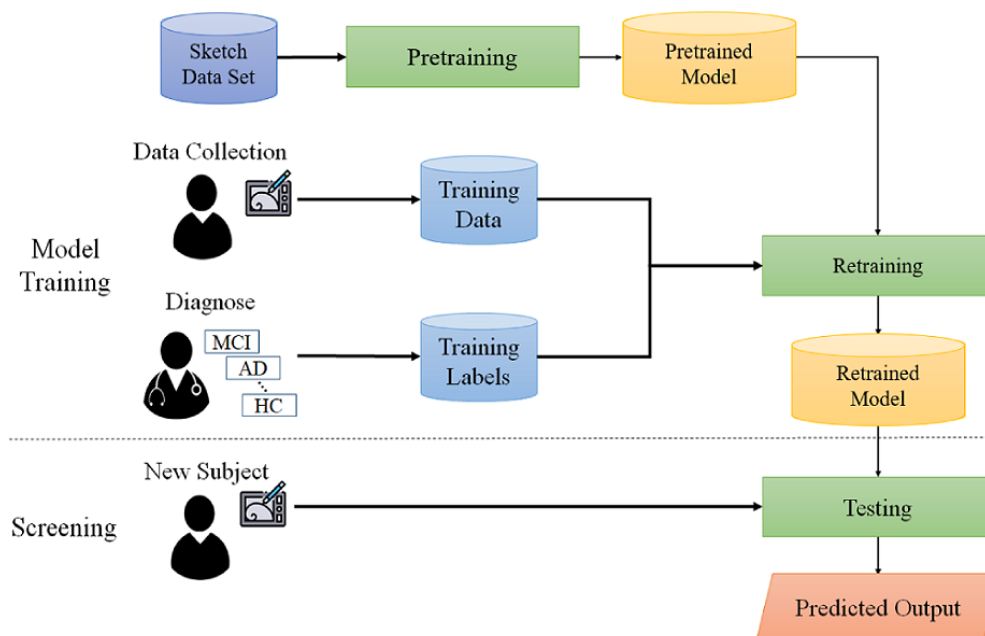
Screening System

Neuropsychological Test Selection

Neuropsychological test selection was based on whether it could be performed in a clinical setting and whether it had been used in related AD and MCI studies. The ROCF test is a neuropsychological test that has been adopted to assess various cognitive functions, such as visuospatial abilities, visual episodic memory, organization skills, attention, and visuomotor coordination [29]. Visual memory impairment is an early sign of AD [30], and some studies [31,32] have shown that the ROCF test can identify patients with MCI, patients with mild AD, and HCs.

The ROCF does not resemble any existing object; it combines many shapes that include lines, circles, rectangles, triangles, crosses, diamonds, and more. There are three trials during an ROCF test: copy, immediate recall, and delayed recall. Cognitive functions such as attention, visuospatial processing, and visuomotor coordination are required for copying the complicated geometrical figures successfully. The immediate recall and the delayed recall are used to assess the participant's ability to retrieve learned information from memory incidentally.

Data Collection Procedure

Overview

First, participants were invited to participate in the study according to ethical approval from the Institutional Review Board (IRB) of the National Taiwan University Hospital (NTUH; see Ethics Approval section for details), and written

informed consent was received from each of them. Each participant was asked to sit at a table with pen and paper or with a Cintiq 16 tablet (Wacom) and Pro Pen 2 (Wacom) [33]. Next, the participant was asked to write his or her name or draw some shapes on the digital device using the smart pen in order to become familiar with the devices. After that, the participant was informed about the process of the ROCF test during three trials: the copy trial, the immediate recall trial, and the delayed recall trial.

Copy Trial

The participant was shown the ROCF and asked to duplicate the complicated geometrical figure as close as possible to the original figure. The participant was informed that there was no time limit for copying the figure. After the copy stage was finished, both the original ROCF and the copied figure that was drawn by the participant were removed from sight. Furthermore, the participant was not notified that the figure would need to be drawn again in the subsequent trials.

Immediate Recall Trial

After a short delay, the participant was asked to draw the complicated geometrical figure from memory with as much detail as possible. The participant was informed that there was no time limit. When the immediate recall drawing was finished, the drawn figure was moved away from the participant's sight.

Delayed Recall Trial

After a 20- to 30-minute delay, the participant was asked to redraw the complicated geometrical figure from memory. The participant was informed that there was no time limit.

Data Preparation

Overview

Two different data sets were used to evaluate our proposed AD screening system; they were gathered according to ethical approval from the IRB of the NTUH (see Ethics Approval section). In line with IRB ethical approval, older adults in Taiwan were invited, and their written informed consent was obtained. Participants with a past or current history of the following conditions were excluded from this study:

1. Nonneurodegenerative problems that might affect brain function, such as stroke, epilepsy, and moderate or severe head injury.
2. Severe psychiatric illness, such as depression and autism.
3. Drug abuse.
4. Blindness or severe hearing impairment that would result in participants not being able to take the ROCF neuropsychological test.

The details of both data sets are described in the following sections.

NTUH_ROCF Data Set

This study data set included a total of 118 participants: 59 (50.0%) participants with MCI and 59 (50.0%) HCs. The NTUH_ROCF data set was collected using pen and paper through the data collection procedure described above. All participants underwent a comprehensive neuropsychological assessment, including measurements from five cognitive domains: attention, executive function, visuospatial function, memory, and language. An expert from our team evaluated the assessments; the criteria of classifying patients with MCI was based on the approaches proposed by Jak et al [34].

Table 1 shows the demographic information of the older adult participants, including gender, age, years of education (minimum 6 years), and MMSE scores for the MCI and HC groups. Participants were asked to draw the ROCF pictures during the copy trial, the 3-minute delayed trial (ie, immediate recall), and the 30-minute delayed trial (ie, delayed recall).

Table 1. Demographic information from the NTUH_ROCF[a] data set.

Characteristic	Participants with mild cognitive impairment (n=59)	Healthy controls (n=59)
Gender, n (%)		
Female	31 (53)	33 (56)
Male	28 (47)	26 (44)
Age (years), mean (SD)	67.51 (6.30)	62.58 (5.89)
Education (years), mean (SD)	13.12 (3.20)	15.05 (2.82)
MMSE[b] score, mean (SD)	27.81 (2.10)	29.18 (0.96)

[a]NTUH_ROCF: National Taiwan University Hospital_Rey-Osterrieth Complex Figure.

[b]MMSE: Mini-Mental State Examination; total scores range from 0 (all answers are incorrect) to 30 (all answers are correct).

NTUH_D-ROCF Data Set

This study data set included a total of 60 participants: 30 (50%) participants with AD and 30 (50%) HCs. Patients with AD were recruited from NTUH, and the NTUH_D-ROCF data set (where "D" represents Alzheimer disease) was collected using the graphics tablet and smart pen (Cintiq 16 and Pro Pen 2; Wacom) to evaluate the automation approach's performance. Disease diagnoses from a board-certified neurologist and a board-certified clinical neuropsychologist were used as the ground truth for training the system. The demographic information from the participants is summarized in Table 2. Participants were asked to draw the ROCF pictures during the copy trial, the immediate recall trial, and the 10-minute delayed recall trial.

Table 2. Demographic information from the NTUH_D-ROCF[a] data set.

Characteristic	Participants with Alzheimer disease (n=30)	Healthy controls (n=30)
Gender, n (%)		
Female	20 (67)	19 (63)
Male	10 (33)	11 (37)
Age (years), mean (SD)	77.67 (6.96)	73.40 (7.24)
Education (years), mean (SD)	11.83 (3.55)	15.03 (2.66)
MMSE[b] score, mean (SD)	21.33 (2.80)	28.50 (1.55)

[a]NTUH_D-ROCF: National Taiwan University Hospital_Alzheimer Disease_Rey-Osterrieth Complex Figure.

[b]MMSE: Mini-Mental State Examination; total scores range from 0 (all answers are incorrect) to 30 (all answers are correct).

XSL•FO
RenderX

Designed Architecture of the Neural Network

Overview

Training a deep CNN from scratch is a time-consuming task and usually requires a large amount of data to achieve the goal of generalization. Generally, it is hard for researchers to collect enough labeled images for each specific task. According to research by Yosinski et al [35], the transfer learning technique applied to deep neural networks could achieve surprising results. They found that initializing the weights of a network by transferring features from almost any number of layers of a pretrained network can retain the generalization ability, even fine-tuning the weights according to the target data set. It

inspired us to use the TU (Technical University)-Berlin sketch data set [36] to pretrain our neural network. The data set consists of 250 different object categories, such as animal, insect, plant, food, furniture, transportation, and instrument, where each category contains 80 sketch images. The data set contains a total of 20,000 hand-drawn sketches. We used that data set because it is large and similar to our collected data, in that both sets of images are sketched and contain the shapes of circles, squares, and lines. The learned weights or pretrained models were then transferred to the target screening engine rather than training the target neural network from scratch. The network structure of our proposed screening system is depicted in Figure 3.

Figure 3. The network structure of the screening system for Alzheimer disease. ROCF: Rey-Osterrieth Complex Figure.
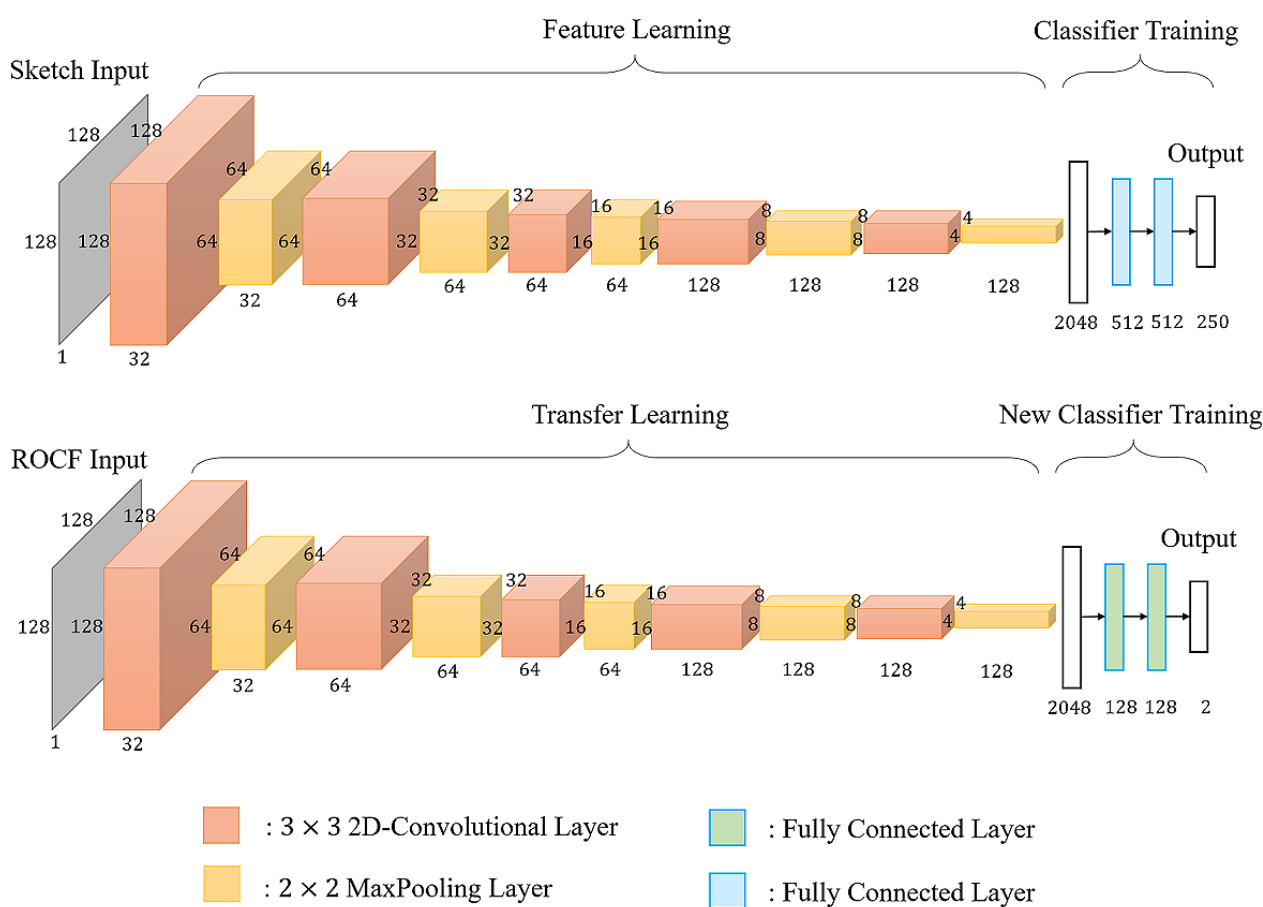
Pretraining Engine

Inspired by the neural network architecture described in Yu et al [37], we further designed a low-cost neural network for pretraining the sketch data set, as demonstrated in the upper part of Figure 3. The input was the image $I \in R^{h \times w \times c}$, where h and w stand for the height and width of the image, and c is the number of channels of the image. The output comprised the probabilities of belonging to the corresponding 250 categories, and the highest probability indicated the predicted output label of the sketch image. In other words, features were extracted based on CNN architecture to recognize the hand-drawn

sketches across numerous object categories. A series of convolutional layers and their following pooling layers acted as the feature extraction, while fully connected (FC) layers were used for further classification.

The pretraining network adopted five 3×3 convolutional layers with a stride of 1 pixel. The stride was set to 1 to keep as much information as possible through the convolution operation. In addition, each convolutional layer was followed by a 2×2 max pooling layer using a stride of 2 pixels. The convolutional function used in each convolutional layer is represented as follows:

where $F^{(l-1)}$ indicates the input feature map to the l-th layer, W is the weight matrix to be applied to the input feature map, b is the bias vector, the operator * is the convolution operation, σ is the nonlinear activation function, *pool* is a subsampling operation, and s represents the pooling size of the filter that usually covers an $s \times s$ square region.

The feature representations were then flattened into a 2048-dimension vector and connected to two FC hidden layers, each with 512 neurons. A rectified linear unit (ReLU) [38] function, as shown in equation 2 below, was used as the activation function of the convolutional layers and the two FC hidden layers, while the softmax function, as shown in equation 3 below, was applied to the output layer to compute the prediction probability for each class. Finally, dropout [39] was adopted after the flattened layer and two FC hidden layers with

a dropout rate of 0.5. The dropout technique was used to prevent overfitting of the training data, which reduced the number of active neurons during training by dropping 50% of the neurons.

A data augmentation technique was used to increase the diversity of sketches per category for classification. Furthermore, it increased the number of training samples through several random transformations on the image, such as vertical shift, horizontal shift, rotation, and flip, in order to train the model with a greater range of various augmented data. This technique lets the model constantly train on new, slightly modified versions of the input data, which enables the model to learn more robust features and increases the generalization of the model. Thus, the shift and rotation transformations of data augmentation were adopted in the training process, and the transformations were then applied in real time as batches were passed into training in this work, as shown in Figure 4.

Figure 4. The real-time process of data augmentation.
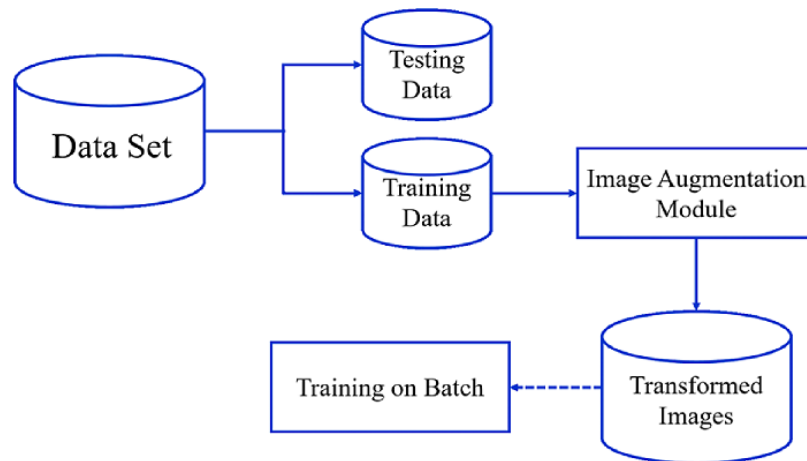
Initially, the sketch data set was separated into training and testing data, and the data augmentation technique was only applied to the training data. The original batch of sketch images was then fed into the image augmentation module to apply a series of random transformations to each image in the batch. Next, the sketches from the training data were randomly shifted horizontally or vertically with a 0.1 fraction of total width or height and randomly rotated in the range of 0.1 degrees. Finally, the new and randomly transformed batch was used for training the CNN, while the original data were not used for training. In other words, the image augmentation module randomly transformed the original images and returned only the new transformed images.

The cross-entropy loss function was applied to calculate the model loss through the training data. We obtained the loss value for later optimization by comparing the model's predictions with the ground truth. The probability denotes the prediction result of i-th class of a sample, where s is the output score of the model, s_i is the i-th element of vector s, and C is the total number of the classes. Set $y = [y_1,...,y_M,...,y_C]$, where $y_M = 1$ and

$y_i = 0$ (if $i \neq M$) to indicate that the M-th class is the ground truth. Then, the cross-entropy loss function L is represented as follows:

Lastly, an Adam optimizer [40] with a learning rate of 0.0001 was used to adjust the trainable parameters to reduce the model loss for each batch. The Adam optimizer combines two methods: AdaGrad (adaptive gradient algorithm) [41], which deals with sparse gradients very well, and RMSProp (root mean square propagation), which does well with online and nonstationary settings.

Retraining Engine

Given the image as the input $I \in R^{h \times w \times c}$, where h and w stand for the height and width of the image and c is the number of channels of the image, which implicitly contains the necessary information for building the screening model of AD, the output is the score or probability of having AD or MCI. A CNN was used to determine the score, and features were learned automatically rather than handcrafted. We formulated the score

of having AD or MCI with a function of the image drawn by the participant, as shown in the following equation:

where *I* represents the image drawn by the participant, the output of the model *score* is a 2D vector, and each dimension is a scalar value between 0 and 1, which indicates the possibility of having AD or MCI or of being an HC, respectively.

The architecture of the retraining engine is shown in the bottom part of Figure 3. The convolutional base of the retraining engine was leveraged from the convolutional base of the pretraining model by using the TU-Berlin sketch data set. The convolutional base layers were frozen, consisting of five convolutional layers with a filter size of 3×3 and a stride of 1 pixel; a 2×2 max pooling layer follows each convolutional layer with a stride of 2 pixels. A new classifier was implemented for further distinguishing the image drawn by the participant. The feature representations were then flattened into a 2048-dimension vector and connected to two FC layers, each with 128 neurons. Every node of the FC layer applied the ReLU [38] activation function (equation 2). The probability or score of having AD or MCI was calculated by applying the softmax function (equation 3). The dropout technique was also applied after the flattened layer with a dropout rate of 0.5.

The same data augmentation technique applied in the pretraining engine was also implemented in the retraining engine. The ROCF training data were randomly shifted horizontally or vertically with a 0.05 fraction of total width or height and randomly rotated in the range of 0.1 degrees. As mentioned before, the real-time data augmentation technique implemented in the screening engine was similar to that implemented in the pretraining engine.

The corresponding ground truth label for the output is as follows: 0 indicates that the participant is healthy, while 1 indicates that the participant has AD or MCI. The loss function is defined as the cross-entropy sum between the predicted output and the ground truth as follows:

where *L* is the loss function, y_i is the ground truth of class *i*, and *y* and *ŷ* are the truth label and output of the screening engine, respectively. In addition, the Adam optimizer with a learning rate of 0.0001 was adopted for training the retraining engine, and the batch size was 16 for the training process.

Ethics Approval

Ethical approval was obtained from the IRB (No. NTUH 201802091RIND) of the NTUH.

Results

Performance Metrics

The performance of the system was measured using four metrics: sensitivity, specificity, accuracy, and area under the receiver operating characteristic curve (AUROC). Sensitivity represents the proportion of actual patients with AD or MCI who are

identified correctly. Specificity denotes the proportion of people who are genuinely healthy older adults who are identified correctly. Accuracy indicates the ratio of correctly classified patients with MCI or AD and cognitively unimpaired older adults to total participants. The receiver operating characteristic (ROC) curve illustrates the relationship between sensitivity and specificity for a given classification model and several given thresholds. If the ROC curve is almost a straight line through the diagonal, it indicates poor performance. When comparing different classification models, the ROC curve of each model can be drawn, and the AUROC is used as an indicator to illustrate the model's performance. Equations for calculating sensitivity and specificity are as follows:

where *TP* (true positive) and *TN* (true negative) denote the number of correct classifications, and where *FP* (false positive) and *FN* (false negative) denote the number of the incorrect classifications.

Evaluation Procedure

A series of experiments were conducted to examine the efficiency of our proposed screening engine. First, the images were resized to 128×128×1, and the data were randomly shuffled to ensure that they were thoroughly mixed. Next, training and testing were executed on a GeForce GTX 1080 Ti GPU (NVIDIA) to evaluate the performance of the implemented classifier through a 10-fold cross-validation procedure. The data set was randomly shuffled to 10 subsets, which were used as testing data in turn, and the other nine subsets were used as training data for each fold test. The 10-fold cross-validation was repeated five times, and each performance score was recorded.

Evaluation of the NTUH_ROCF Data Set

Comparison of Different ROCF Trials

The performance of the copy, immediate recall, and delayed recall trials were calculated separately, and the results are listed in Table 3. The performance of the copy trial had a mean sensitivity of 0.668 (SD0.015), a mean specificity of 0.536 (SD0.026), a mean accuracy of 0.602 (SD0.009), and a mean AUROC of 0.672 (SD0.004). The results of the copy trial indicate that it was not easy to distinguish whether a participant had MCI or was an HC; this might be the case because both patients with MCI and HCs still have adequate attention and visuospatial processing ability, allowing them to duplicate the complex geometrical figure during the copy trial. On the contrary, the delayed recall trial had the best classification capability in differentiating participants with MCI from HCs, with a mean sensitivity of 0.847 (SD0.017), a mean specificity of 0.905 (SD0.009), a mean accuracy of 0.876 (SD0.010), and a mean AUROC of 0.913 (SD0.004). The performance of the immediate recall trial had a mean sensitivity of 0.736 (SD0.028), a mean specificity of 0.885 (SD0.014), a mean accuracy of 0.810 (SD0.015), and a mean AUROC of 0.871 (SD0.008). Compared with cognitively unimpaired older adults, the patients with MCI may have had problems recalling the figure from memory after some time.

Table 3. Performance of three ROCF[a] trials using the NTUH_ROCF[b] data set.

Metric	Copy trial, mean (SD)	Immediate recall trial, mean (SD)	Delayed recall trial, mean (SD)
Sensitivity	0.668 (0.015)	0.736 (0.028)	0.847(0.017)
Specificity	0.536 (0.026)	0.885 (0.014)	0.905 (0.009)
Accuracy	0.602 (0.009)	0.810 (0.015)	0.876 (0.010)
AUROC[c]	0.672 (0.004)	0.871 (0.008)	0.913 (0.004)

[a]ROCF: Rey-Osterrieth Complex Figure.

[b]NTUH_ROCF: National Taiwan University Hospital_Rey-Osterrieth Complex Figure.

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

Performance of the Proposed Screening System for Classifying Participants With MCI Versus Healthy Controls

In this experiment, the performance of the proposed architecture of the screening engine for distinguishing between the complex figures drawn by participants with MCI and HCs was evaluated using the images drawn by the participants during the delayed recall trial. Initially, the TU-Berlin sketch data set was used to pretrain the neural network; the learned feature representations were then leveraged for our ROCF data set for further training. Figure 5 shows the mean (SD) of AUROC and accuracy for each repeat of the 10-fold cross-validation and the mean (SD) of these five repeats. The performance of our model achieved a mean AUROC of 0.913 (SD 0.004), while the mean accuracy of the five repeats of 10-fold cross-validation was 0.876 (SD 0.010).

Figure 5. Receiver operating characteristic (ROC) curves of the proposed screening engine after five repeats of 10-fold cross-validation using the NTUH_ROCF data set. Acc: accuracy; AUROC: area under the receiver operating characteristic curve; NTUH_ROCF: National Taiwan University Hospital_Rey-Osterrieth Complex Figure.
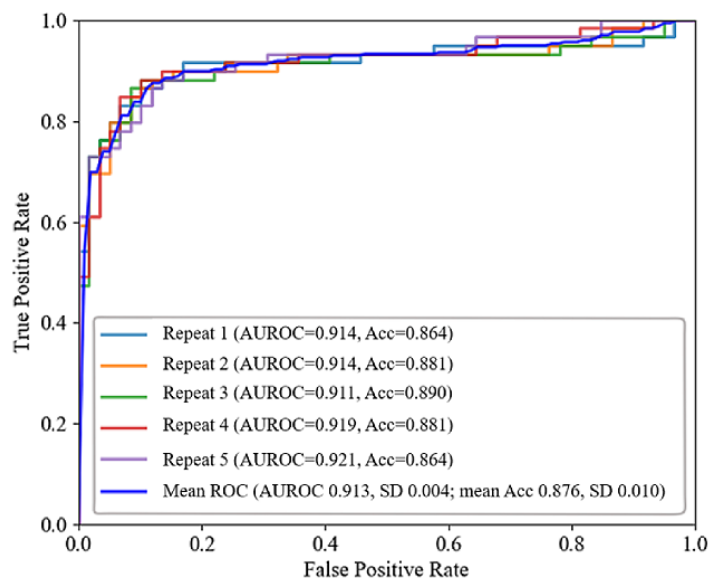
Evaluation of the NTUH_D-ROCF Data Set

Comparison of Different ROCF Trials

Different ROCF trials were evaluated individually, and their performance results are shown in Table 4. The performance of the immediate recall trial and the 10-minute delayed recall trial were similar in distinguishing between participants with AD and HCs. The performance of delayed recall had a mean sensitivity of 0.820 (SD0.038), a mean specificity of 0.953 (SD 0.018), a mean accuracy of 0.887 (SD 0.016), and a mean AUROC of 0.940 (SD 0.006). The performance of immediate recall had a mean sensitivity of 0.827 (SD 0.015), a mean specificity of 0.947 (SD 0.018), a mean accuracy of 0.887 (SD 0.012), and a mean AUROC of 0.950 (SD 0.003). The results showed that the immediate recall trial had the best performance, followed by the 10-minute delayed recall trial, while both could be used to distinguish between participants with AD and HCs. The patients with AD may have had problems recalling the complex figure from memory during the immediate recall trial and the 10-minute delayed recall trial, as compared to HCs. On the other hand, compared with the immediate recall trial or the 10-minute delayed recall trial, the performance of the copy trial was less discriminative; this trial had a mean sensitivity of 0.627 (SD 0.028), a mean specificity of 0.900 (SD 0.033), a mean accuracy of 0.763 (SD 0.016), and a mean AUROC of 0.762 (SD 0.018).

Table 4. Performance of three ROCF[a] trials using the NTUH_D-ROCF[b] data set.

Metric	Copy trial, mean (SD)	Immediate recall trial, mean (SD)	Delayed recall trial, mean (SD)
Sensitivity	0.627 (0.028)	0.827 (0.015)	0.820 (0.038)
Specificity	0.900 (0.033)	0.947 (0.018)	0.953 (0.018)
Accuracy	0.763 (0.016)	0.887 (0.012)	0.887 (0.016)
AUROC[c]	0.762 (0.018)	0.950 (0.003)	0.940 (0.006)

[a]ROCF: Rey-Osterrieth Complex Figure.

[b]NTUH_D-ROCF: National Taiwan University Hospital_Alzheimer Disease_Rey-Osterrieth Complex Figure.

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

Performance of the Proposed Screening System for Classifying Participants With AD Versus Healthy Controls

The performance of the proposed architecture of the screening engine to distinguish between the abstract and complex figures drawn by participants with AD and HCs was conducted using the images collected from the immediate recall trial. First, the TU-Berlin sketch data set was also used to pretrain the neural network; the feature representations learned by the pretrained neural network were then fine-tuned and leveraged for our ROCF data set for further training. As a result, the performance of our model achieved a mean AUROC of 0.950 (SD 0.003), while the mean accuracy of the five repeats of 10-fold cross-validation was 0.887 (SD 0.012), as shown in Figure 6.

Figure 6. Receiver operating characteristic (ROC) curves of the proposed screening engine after five repeats of 10-fold cross-validation using the NTUH_D-ROCF data set. Acc: accuracy; AUROC: area under the receiver operating characteristic curve; NTUH_D-ROCF: National Taiwan University Hospital_Alzheimer Disease_Rey-Osterrieth Complex Figure.
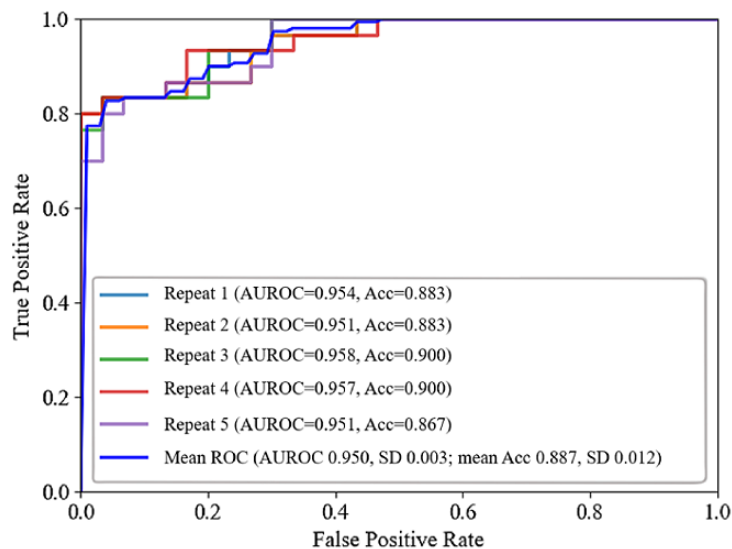
Effectiveness of the Dropout and Data Augmentation Techniques

An experiment verifying the performance of the designed system after applying the data augmentation and dropout techniques was conducted for the data collected in the delayed recall trial. The data augmentation method was adopted when training the neural network concurrently; only the training data, instead of testing data, were augmented. The performance of the designed system after applying both techniques was better, with a mean sensitivity of 0.847 (SD 0.017), a mean specificity of 0.905 (SD 0.009), a mean accuracy of 0.876 (SD 0.010), and a mean AUROC of 0.913 (SD 0.004). When the techniques were not used, the system had a mean sensitivity of 0.824 (SD 0.019), a mean specificity of 0.898 (SD0.024), a mean accuracy of 0.861 (SD 0.001), and a mean AUROC of 0.893 (SD 0.012), as seen in Table 5. When applying data augmentation and dropout techniques, most studies on image classification obtain better results. Data augmentation techniques could extend the diversity of the training data, and dropout techniques could avoid coadaptation of the model by randomly disabling neurons with probability during the training process. The results showed that with the data augmentation and dropout techniques, the system performed better, according to the results provided in Table 5. Therefore, integrating them into the model provides better results. Consequently, to obtain better performance, both technologies were adopted in our model. The higher the sensitivity, specificity, accuracy, and AUROC values, the better the performance was. However, these numbers do not explain the system's reliability, sustainability, and consistency. That will be a different concern to address, which is out of the scope of this research.

Table 5. Effects of data augmentation and dropout techniques applied to the NTUH_ROCF[a] data set.

Metric	Delayed recall trial, mean (SD)	
	Without data augmentation and dropout techniques	With data augmentation and dropout techniques
Sensitivity	0.824 (0.019)	0.847 (0.017)
Specificity	0.898 (0.024)	0.905 (0.009)
Accuracy	0.861 (0.011)	0.876 (0.010)
AUROC[b]	0.893 (0.012)	0.913 (0.004)

[a]NTUH_ROCF: National Taiwan University Hospital_Rey-Osterrieth Complex Figure.

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

Comparison of Different Network Architectures

From the images drawn by participants in the delayed recall trial, the performances of the different architectures of the neural network classifier were studied. Additionally, the total number of parameters and the time to complete 1-fold training were listed for comparison for different classifiers. The different neural network architectures included AlexNet [42]; Sketch-a-Net [37]; our previous work, a convolutional autoencoder neural network [43]; and the proposed network architectures mentioned in this study. As a result, the architecture of our proposed framework in this study achieved better performance than the architecture of AlexNet, Sketch-a-Net, and our previous work, as shown in Table 6.

Table 6. Performance of different network architectures applied to the NTUH_ROCF[a] data set.

Metric	AlexNet	Sketch-a-Net	Our system	
			Without data augmentation and dropout techniques	With data augmentation and dropout techniques
Sensitivity, mean (SD)	0.698 (0.039)	0.671 (0.047)	0.756 (0.033)	0.847 (0.017)
Specificity, mean (SD)	0.790 (0.046)	0.820 (0.054)	0.864 (0.017)	0.905 (0.009)
Accuracy, mean (SD)	0.744 (0.034)	0.746 (0.019)	0.810 (0.020)	0.876 (0.010)
AUROC,[b] mean (SD)	0.814 (0.021)	0.819 (0.009)	0.851 (0.020)	0.913 (0.004)
Total parameters ($\times 10^6$), n	46.73	8.38	0.40	0.56
Time required to complete 1-fold training, minutes	10	29	9	2

[a]NTUH_ROCF: National Taiwan University Hospital_Rey-Osterrieth Complex Figure.

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

The sensitivity, specificity, accuracy, and AUROC of our proposed network architecture achieved the highest performance compared to the others mentioned above. Furthermore, the total number of parameters used in our proposed model was 560,000, which was relatively fewer parameters than that used with AlexNet (83.45 times larger) and Sketch-a-Net (14.96 times larger). Although the total number of parameters in our previous work was 0.4 million (1.4 times smaller), the accuracy and AUROC of our proposed model increased by 6.6% and 6.2%, respectively. Moreover, it took only 2 minutes to complete 1-fold training of our proposed network architecture compared to AlexNet (10 minutes), Sketch-a-Net (29 minutes), and our previous work (9 minutes). Figure 7 depicts the ROC curves of the different classifiers.

Figure 7. Receiver operating characteristic curves of the different network architectures using the NTUH_ROCF data set. Acc: accuracy; AUROC: area under the receiver operating characteristic curve; NTUH_ROCF: National Taiwan University Hospital_Rey-Osterrieth Complex Figure.
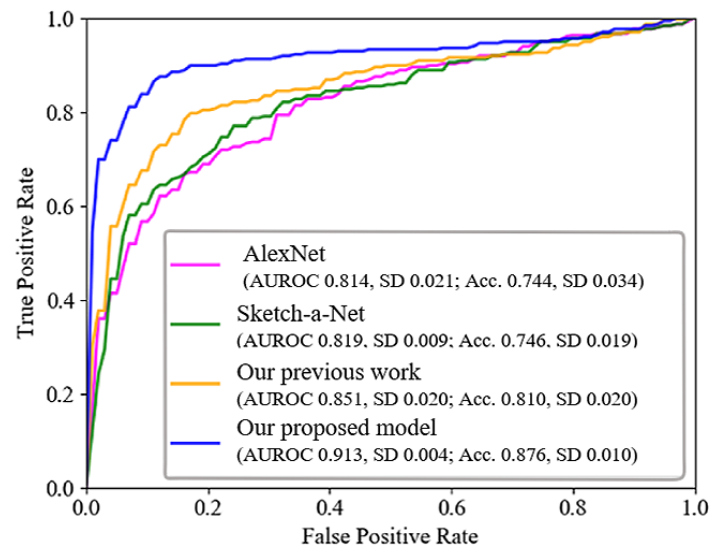
Effectiveness of the Transfer Learning Technique

A comparison of models with or without use of the transfer learning strategy was carried out using the images drawn by the participants in the delayed recall trial. In order to validate the effectiveness of the transfer learning method, the network applied the same structure as that of the convolutional base of the pretraining model using the TU-Berlin sketch data set mentioned in the Methods section. The network was composed of five 3×3 convolutional layers; a 2×2 max pooling layer followed each convolution layer, and two FC layers with neurons were used to discriminate the images drawn by the participants. Moreover, the dropout and data augmentation techniques were also implemented. In addition, we compared the transfer learning technique using Sketch-a-Net's network architecture [37]. First, the network architecture of Sketch-a-Net

was pretrained using the TU-Berlin data set [36]. The pretrained model was then transferred to our NTUH_ROCF data set, and data augmentation was also implemented for further training and classification of participants with MCI or HCs.

As a result, the transfer learning technique, which pretrained using a larger data set, achieved better performance, with a mean sensitivity of 0.847 (SD 0.017), a mean specificity of 0.905 (SD 0.009), a mean accuracy of 0.876 (SD 0.010), and a mean AUROC of 0.913 (SD 0.004). When the transfer learning technique was not used, the model performance achieved a mean sensitivity of 0.749 (SD 0.030), a mean specificity of 0.814 (SD 0.017), a mean accuracy of 0.781 (SD 0.014), and a mean AUROC of 0.846 (SD 0.005), as shown in Table 7. Moreover, our proposed network achieved better results than the Sketch-a-Net with the transfer learning architecture.

Table 7. Performance of network architectures with and without transfer learning applied to the NTUH_ROCF[a] data set.

| Metric	Sketch-a-Net: with transfer learning, mean (SD)	Our proposed model, mean (SD)	
		Without transfer learning	With transfer learning
Sensitivity	0.641 (0.040)	0.749 (0.030)	0.847 (0.017)
Specificity	0.810 (0.022)	0.814 (0.017)	0.905 (0.009)
Accuracy	0.725 (0.010)	0.781 (0.014)	0.876 (0.010)
AUROC[b]	0.819 (0.010)	0.846 (0.005)	0.913 (0.004)

[a]NTUH_ROCF: National Taiwan University Hospital_Rey-Osterrieth Complex Figure.

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

Discussion

System Usage

The developed system is applicable for use as an early-stage screening system in hospitals. It could help clinicians diagnose patients with MCI and AD. It could also help clinicians assess patients' visual perception and their ability to retrieve learned information, in order to test their long-term visual memory

function. The accuracy, reliability, and efficiency of the screening system is important for diagnosing patients correctly.

Limitations of This Study

For the data set, as ground truth, it is assumed that the participants were diagnosed correctly by experienced doctors and neuropsychologists. To study the designed system that has been proposed, only the characteristics that are detectable from

the neuropsychology test were involved. Therefore, it is a challenge to have participants participate in the study. For research purposes, the number of data sets obtained was minimal, and the data set was only collected locally in Taiwan. Therefore, the data set is biased. As for this study's research purpose, the study was focused on distinguishing participants with AD and MCI from HCs in an Asian older adult population. In order to obtain more generalized data sets to reduce overfitting, further data need to be collected from participants of different ethnicities and age groups. This system is only useful for one specific neuropsychological test: the ROFC test. In the future, incorporation with other neuropsychological tests will improve the performance of the screening system.

Conclusions

For decades, AD has been one of the most common diseases among older adults. It is challenging to identify the difference in cognitive performance between patients with MCI and people experiencing normal aging, as the difference may be very subtle, particularly at the early stage of MCI. Nevertheless, early identification of individuals with a high risk of developing AD will help in the management and support of the long-term quality of life of patients with AD and their caregivers. Neuropsychology and cognitive ability can be tested during the screening process, and they do not require any sophisticated medical equipment. Among different types of cognitive testing, clinicians and neuropsychologists often use the ROCF test to help with diagnosing patients. However, it involves intensive labor, and the tester must be qualified as an expert. Data-driven

deep learning approaches, which can extract features automatically, have opened the door to the possibility of assisting clinicians, such as neurologists, and clinical neuropsychologists during screening by making the diagnosis process more effective than the traditional approach. With the aid of transfer learning and deep learning, we have proposed an automatic digital screening system to characterize hand-drawn images. It allows us to effectively distinguish patients with MCI and AD from people experiencing normal aging based on the ROCF test process.

The digital screening system that was developed in this study has shown promising preliminary results regarding distinguishing patients with AD and MCI from HCs. Therefore, this screening system can be used during early assessments to diagnose individuals with a high risk of AD. The results have also shown that the system performed better when distinguishing patients with AD from HCs, since there is a significant characteristic difference, as compared to distinguishing patients with MCI from HCs. After analyzing the drawn images, the scores were calculated automatically, and the calculation time was swift. Therefore, this system can replace the labor-intensive and time-consuming work that comes with manually calculating scores according to the criteria of the scoring system. For future studies, merging additional data from various types and stages of dementia will increase the capability of our system in assisting clinicians. Moreover, other types of neuropsychological tests can be included through ensemble methods to provide a complete screening system.

Conflicts of Interest

None declared.

References

1. World Alzheimer Report 2019: Attitudes to Dementia. London, UK: Alzheimer's Disease International; 2019 Sep. URL: https://www.alzint.org/u/WorldAlzheimerReport2019.pdf [accessed 2022-02-28]
2. The top 10 causes of death. World Health Organization. 2020 Dec 09. URL: https://www.who.int/news-room/fact-sheets/detail/the-top-10-causes-of-death [accessed 2019-10-08]
3. Global Action Plan on the Public Health Response to Dementia: 2017 - 2025. Geneva, Switzerland: World Health Organization; 2017 Dec 07. URL: https://apps.who.int/iris/bitstream/handle/10665/259615/9789241513487-eng.pdf?sequence=1 [accessed 2022-02-28]
4. Galvin JE, Sadowsky CH, NINCDS-ADRDA. Practical guidelines for the recognition and diagnosis of dementia. J Am Board Fam Med 2012;25(3):367-382 [FREE Full text] [doi: 10.3122/jabfm.2012.03.100181] [Medline: 22570400]
5. Sperling RA, Aisen PS, Beckett LA, Bennett DA, Craft S, Fagan AM, et al. Toward defining the preclinical stages of Alzheimer's disease: Recommendations from the National Institute on Aging-Alzheimer's Association workgroups on diagnostic guidelines for Alzheimer's disease. Alzheimers Dement 2011 May;7(3):280-292 [FREE Full text] [doi: 10.1016/j.jalz.2011.03.003] [Medline: 21514248]
6. Petersen RC, Lopez O, Armstrong MJ, Getchius TSD, Ganguli M, Gloss D, et al. Practice guideline update summary: Mild cognitive impairment: Report of the Guideline Development, Dissemination, and Implementation Subcommittee of the American Academy of Neurology. Neurology 2018 Jan 16;90(3):126-135. [doi: 10.1212/WNL.0000000000004826] [Medline: 29282327]

7.   Folstein MF, Folstein SE, McHugh PR. "Mini-mental state". A practical method for grading the cognitive state of patients for the clinician. J Psychiatr Res 1975 Nov;12(3):189-198. [Medline: 1202204]

8.   Hughes CP, Berg L, Danziger WL, Coben LA, Martin RL. A new clinical scale for the staging of dementia. Br J Psychiatry 1982 Jun;140:566-572. [Medline: 7104545]

9.   Dubois B, Hampel H, Feldman HH, Scheltens P, Aisen P, Andrieu S, et al. Preclinical Alzheimer's disease: Definition, natural history, and diagnostic criteria. In: Proceedings of the Meeting of the International Working Group (IWG) and the American Alzheimer's Association on "The Preclinical State of AD". 2016 Presented at: Meeting of the International Working Group (IWG) and the American Alzheimer's Association on "The Preclinical State of AD"; July 23, 2015; Washington, DC p. 292-323 URL: http://europepmc.org/abstract/MED/27012484 [doi: 10.1016/j.jalz.2016.02.002]

10.   Bogdanovic N. The challenges of diagnosis in Alzheimer's disease. US Neurol 2018;14:15. [doi: 10.17925/usn.2018.14.1.15]

11.   Pennanen C, Kivipelto M, Tuomainen S, Hartikainen P, Hänninen T, Laakso MP, et al. Hippocampus and entorhinal cortex in mild cognitive impairment and early AD. Neurobiol Aging 2004 Mar;25(3):303-310. [doi: 10.1016/S0197-4580(03)00084-8] [Medline: 15123335]

12.   Dickerson BC, Sperling RA. Functional abnormalities of the medial temporal lobe memory system in mild cognitive impairment and Alzheimer's disease: Insights from functional MRI studies. Neuropsychologia 2008;46(6):1624-1635 [FREE Full text] [doi: 10.1016/j.neuropsychologia.2007.11.030] [Medline: 18206188]

13.   Hu K, Wang Y, Chen K, Hou L, Zhang X. Multi-scale features extraction from baseline structure MRI for MCI patient classification and AD early diagnosis. Neurocomputing 2016 Jan;175:132-145. [doi: 10.1016/j.neucom.2015.10.043]

14.   Challis E, Hurley P, Serra L, Bozzali M, Oliver S, Cercignani M. Gaussian process classification of Alzheimer's disease and mild cognitive impairment from resting-state fMRI. Neuroimage 2015 May 15;112:232-243 [FREE Full text] [doi: 10.1016/j.neuroimage.2015.02.037] [Medline: 25731993]

15.   Li F, Tran L, Thung K, Ji S, Shen D, Li J. A robust deep model for improved classification of AD/MCI patients. IEEE J Biomed Health Inform 2015 Sep;19(5):1610-1616 [FREE Full text] [doi: 10.1109/JBHI.2015.2429556] [Medline: 25955998]

16.   Casaletto KB, Heaton RK. Neuropsychological assessment: Past and future. J Int Neuropsychol Soc 2017 Dec 04;23(9-10):778-790. [doi: 10.1017/s1355617717001060]

17.   Breton A, Casey D, Arnaoutoglou NA. Cognitive tests for the detection of mild cognitive impairment (MCI), the prodromal stage of dementia: Meta-analysis of diagnostic accuracy studies. Int J Geriatr Psychiatry 2019 Feb;34(2):233-242. [doi: 10.1002/gps.5016] [Medline: 30370616]

18.   Prange A, Sonntag D. Modeling cognitive status through automatic scoring of a digital version of the Clock Drawing Test. In: Proceedings of the 27th ACM Conference on User Modeling, Adaptation and Personalization. 2019 Presented at: 27th ACM Conference on User Modeling, Adaptation and Personalization; June 9-12, 2019; Larnaca, Cyprus p. 70-77. [doi: 10.1145/3320435.3320452]

19.   Mendez MF, Ala T, Underwood KL. Development of scoring criteria for the clock drawing task in Alzheimer's disease. J Am Geriatr Soc 1992 Nov;40(11):1095-1099. [doi: 10.1111/j.1532-5415.1992.tb01796.x] [Medline: 1401692]

20.   Ehreke L, Luppa M, König H, Riedel-Heller SG. Is the Clock Drawing Test a screening tool for the diagnosis of mild cognitive impairment? A systematic review. Int Psychogeriatr 2010 Feb;22(1):56-63. [doi: 10.1017/S1041610209990676] [Medline: 19691908]

21.   Meyers JE, Meyers KR. Rey complex figure test under four different administration procedures. Clin Neuropsychol 1995 Feb;9(1):63-67. [doi: 10.1080/13854049508402059]

22.   Rey A. L'examen psychologique dans les cas d'encéphalopathie traumatique (Les problems). Arch Psychol (Geneve) 1941;28:215-285.

23.   Osterrieth PA. Le test de copie d'une figure complexe; contribution à l'étude de la perception et de la mémoire. Arch Psychol (Geneve) 1944;30:206-356. [doi: 10.3406/bupsy.1949.5612]

24.   Miller J, Hanson E, Baerresen K, Miller K, Gottuso A, Ercoli L, et al. Screening for mild cognitive impairment (MCI) with the Mini-Mental Status Exam (MMSE) and Rey-Osterrieth Complex Figure Test (ROCF). Arch Clin Neuropsychol 2014 Aug 28;29(6):508. [doi: 10.1093/arclin/acu038.12]

25.   Salvadori E, Dieci F, Caffarra P, Pantoni L. Qualitative evaluation of the immediate copy of the Rey-Osterrieth Complex Figure: Comparison between vascular and degenerative MCI patients. Arch Clin Neuropsychol 2019 Feb 01;34(1):14-23. [doi: 10.1093/arclin/acy010] [Medline: 29420698]

26.   Stern R, Singer E, Duke L, Singer N, Morey C, Daughtrey E, et al. The Boston qualitative scoring system for the Rey-Osterrieth complex figure: Description and interrater reliability. Clin Neuropsychol 1994 Aug 18;8(3):309-322. [doi: 10.1080/13854049408404137]

27.   Corwin J, Bylsma FW. Psychological examination of traumatic encephalopathy. Clin Neuropsychol 1993 Jan;7(1):3-21. [doi: 10.1080/13854049308401883]

28.   Bernstein JH, Waber DP. Developmental Scoring System for the Rey-Osterrieth Complex Figure: Professional Manual. Lutz, FL: Psychological Assessment Resources, Inc; 1996. URL: https://www.nctsn.org/measures/developmental-scoring-system-rey-osterrieth-complex [accessed 2022-02-28]

29.   Shin M, Park S, Park S, Seol S, Kwon JS. Clinical and empirical applications of the Rey-Osterrieth Complex Figure Test. Nat Protoc 2006;1(2):892-899. [doi: 10.1038/nprot.2006.115] [Medline: 17406322]

XSL•FO
RenderX

30. Kawas CH, Corrada MM, Brookmeyer R, Morrison A, Resnick SM, Zonderman AB, et al. Visual memory predicts Alzheimer's disease more than a decade before diagnosis. Neurology 2003 Apr 08;60(7):1089-1093. [doi: 10.1212/01.wnl.0000055813.36504.bf] [Medline: 12682311]

31. Alladi S, Arnold R, Mitchell J, Nestor PJ, Hodges JR. Mild cognitive impairment: Applicability of research criteria in a memory clinic and characterization of cognitive profile. Psychol Med 2006 Apr;36(4):507-515. [doi: 10.1017/S0033291705006744] [Medline: 16426486]

32. Kasai M, Meguro K, Hashimoto R, Ishizaki J, Yamadori A, Mori E. Non-verbal learning is impaired in very mild Alzheimer's disease (CDR 0.5): Normative data from the learning version of the Rey-Osterrieth Complex Figure Test. Psychiatry Clin Neurosci 2006 Apr;60(2):139-146 [FREE Full text] [doi: 10.1111/j.1440-1819.2006.01478.x] [Medline: 16594936]

33. Wacom. URL: https://www.wacom.com/en-in/ [accessed 2018-12-08]

34. Jak AJ, Bondi MW, Delano-Wood L, Wierenga C, Corey-Bloom J, Salmon DP, et al. Quantification of five neuropsychological approaches to defining mild cognitive impairment. Am J Geriatr Psychiatry 2009 May;17(5):368-375 [FREE Full text] [doi: 10.1097/JGP.0b013e31819431d5] [Medline: 19390294]

35. Yosinski J, Clune J, Bengio Y, Lipson H. How transferable are features in deep neural networks? In: Proceedings of the 27th International Conference on Neural Information Processing Systems. 2014 Presented at: 27th International Conference on Neural Information Processing Systems; December 8-13, 2014; Montreal, QC p. 3320-3328 URL: https://proceedings.neurips.cc/paper/2014/file/375c71349b295fbe2dcdca9206f20a06-Paper.pdf

36. Eitz M, Hays J, Alexa M. How do humans sketch objects? ACM Trans Graph 2012 Aug 05;31(4):1-10. [doi: 10.1145/2185520.2185540]

37. Yu Q, Yang Y, Song YZ, Xiang T, Hospedales Y. Sketch-a-Net that beats humans. In: Proceedings of the British Machine Vision Conference. 2015 Sep Presented at: British Machine Vision Conference; September 7-10, 2015; Swansea, UK p. 7.1-7.12. [doi: 10.5244/c.29.7]

38. Nair V, Hinton GE. Rectified linear units improve restricted Boltzmann machines. In: Proceedings of the 27th International Conference on Machine Learning. 2010 Presented at: 27th International Conference on Machine Learning; June 21-24, 2010; Haifa, Israel p. 807-814 URL: https://www.cs.toronto.edu/~fritz/absps/reluICML.pdf

39. Srivastava N. Improving Neural Networks With Dropout [master's thesis]. Toronto, ON: University of Toronto; 2013. URL: http://www.cs.toronto.edu/~nitish/msc_thesis.pdf [accessed 2022-02-28]

40. Kingma DP, Ba LL. Adam: A method for stochastic optimization. In: Proceedings of the 3rd International Conference on Learning Representations. 2015 Presented at: The 3rd International Conference on Learning Representations; May 7-9, 2015; San Diego, CA URL: https://arxiv.org/pdf/1412.6980.pdf

41. Duchi J, Hazan E, Singer Y. Adaptive subgradient methods for online learning and stochastic optimization. J Mach Learn Res 2011 Jul;12:2121-2159 [FREE Full text]

42. Krizhevsky A, Sutskever I, Hinton GE. ImageNet classification with deep convolutional neural networks. Commun ACM 2017 May 24;60(6):84-90. [doi: 10.1145/3065386]

43. Cheah W, Chang W, Hwang J, Hong S, Fu L, Chang Y. A screening system for mild cognitive impairment based on neuropsychological drawing test and neural network. In: Proceedings of the IEEE International Conference on Systems, Man and Cybernetics. 2019 Presented at: IEEE International Conference on Systems, Man and Cybernetics; October 6-9, 2019; Bari, Italy p. 3543-3548. [doi: 10.1109/smc.2019.8913880]

Abbreviations

AD: Alzheimer disease
AdaGrad: adaptive gradient algorithm
AUROC: area under the receiver operating characteristic curve
BQSS: Boston Qualitative Scoring System
CDT: Clock-Drawing Test
CNN: convolutional neural network
D: Alzheimer disease, in the context of NTUH_D-ROCF
FC: fully connected
FN: false negative
FP: false positive
HC: healthy control
IRB: Institutional Review Board
MCI: mild cognitive impairment
MMSE: Mini-Mental State Examination
MRI: magnetic resonance imaging
NTUH: National Taiwan University Hospital
PET: positron emission tomography
ReLU: rectified linear unit

RMSProp: root mean square propagation
ROC: receiver operating characteristic
ROCF: Rey-Osterrieth Complex Figure
SVM: support vector machine
TN: true negative
TP: true positive
TU: Technical University

<u>Original Paper</u>

Pandemic-Related Impairment in the Monitoring of Patients With Hypertension and Diabetes and the Development of a Digital Solution for the Community Health Worker: Quasiexperimental and Implementation Study

Christiane Correa Rodrigues Cimini[1], MD, MSc, PhD; Junia Xavier Maia[2], MD; Magda Carvalho Pires[3], BA, MSc, PhD; Leonardo Bonisson Ribeiro[2], BA; Vânia Soares de Oliveira e Almeida Pinto[1], MD, MSc; James Batchelor[4], BA; Antonio Luiz Pinho Ribeiro[5], MD, PhD; Milena Soriano Marcolino[2], MD, MSc, PhD

[1]Medical School and Telehealth Center, Universidade Federal dos Vales do Jequitinhonha e Mucuri, Teófilo Otoni-MG, Brazil

[2]Telehealth Center, Hospital das Clínicas, Faculdade de Medicina, Universidade Federal de Minas Gerais, Belo Horizonte, Brazil

[3]Department of Statistics, Universidade Federal de Minas Gerais, Belo Horizonte, Brazil

[4]Clinical Informatics Research Unit, Faculty of Medicine, University of Southampton, Southampton, United Kingdom

[5]Telehealth Center and Cardiology Service, Hospital das Clínicas, Faculdade de Medicina, Universidade Federal de Minas Gerais, Belo Horizonte, Brazil

Corresponding Author:
Christiane Correa Rodrigues Cimini, MD, MSc, PhD
Medical School and Telehealth Center
Universidade Federal dos Vales do Jequitinhonha e Mucuri
Rua do Cruzeiro, 01
Bairro Jardim São Paulo
Teófilo Otoni-MG, 38803-371
Brazil
Phone: 55 33988900906
Email: christiane.cimini@gmail.com

Abstract

Background: The restrictions imposed by the COVID-19 pandemic reduced health service access by patients with chronic diseases. The discontinuity of care is a cause of great concern, mainly in vulnerable regions.

Objective: This study aimed to assess the impact of the COVID-19 pandemic on people with hypertension and diabetes mellitus (DM) regarding the frequency of consultations and whether their disease was kept under control. The study also aimed to develop and implement a digital solution to improve monitoring at home.

Methods: This is a multimethodological study. A quasiexperimental evaluation assessed the impact of the pandemic on the frequency of consultations and control of patients with hypertension and DM in 34 primary health care centers in 10 municipalities. Then, an implementation study developed an app with a decision support system (DSS) for community health workers (CHWs) to identify and address at-risk patients with uncontrolled hypertension or DM. An expert panel assessment evaluated feasibility, usability, and utility of the software.

Results: Of 5070 patients, 4810 (94.87%) had hypertension, 1371 (27.04%) had DM, and 1111 (21.91%) had both diseases. There was a significant reduction in the weekly number of consultations (107, IQR 60.0-153.0 before vs 20.0, IQR 7.0-29.0 after social restriction; $P<.001$). Only 15.23% (772/5070) of all patients returned for a consultation during the pandemic. Individuals with hypertension had lower systolic (120.0, IQR 120.0-140.0 mm Hg) and diastolic (80.0, IQR 80.0-80.0 mm Hg) blood pressure than those who did not return (130.0, IQR 120.0-140.0 mm Hg and 80.0, IQR 80.0-90.0 mm Hg, respectively; $P<.001$). Also, those who returned had a higher proportion of controlled hypertension (64.3% vs 52.8%). For DM, there were no differences in glycohemoglobin levels. Concerning the DSS, the experts agreed that the CHWs can easily incorporate it into their routines and the app can identify patients at risk and improve treatment.

Conclusions: The COVID-19 pandemic caused a significant drop in the number of consultations for patients with hypertension and DM in primary care. A DSS for CHW has proved to be feasible, useful, and easily incorporated into their routines.

XSL•FO
RenderX

Introduction

The COVID-19 pandemic severely hit health care systems worldwide, challenged their responsiveness, and forced them to redistribute human and material resources to emergency services and intensive care units dedicated to COVID-19 patients [1]. This emergency reorganization had a negative impact on monitoring people with noncommunicable chronic diseases (NCDs), such as hypertension and diabetes mellitus (DM), which need continuous follow-up [2].

Mortality from NCDs in low-income and medium-income countries is high due to the limitations of health systems in providing treatment for these diseases [3]. Hypertension is the main modifiable risk factor, with an independent association with cardiovascular disease, chronic kidney disease, and premature death [4] and risk factors for severe COVID-19 and COVID-19 mortality [5,6]. Social restriction intensified risky behaviors, including increased sedentary lifestyles, time in front of screens [7], ultraprocessed food consumption, and number of cigarettes smoked [8]. These habits contribute to weight gain as well as uncontrolled blood pressure (BP) and glucose levels of individuals with hypertension and DM, respectively [9].

This global impact of COVID-19 is especially favorable for the adoption of digital solutions in response to the challenges that the pandemic has imposed. They can be used not only to follow people with suspected or confirmed of COVID-19 but also to monitor patients with other diseases and provide essential health care services at the community level. Therefore, the pandemic has led to rapid development and utilization of mobile health (mHealth) apps [10,11], although these tools have been available for a long time.

Since June 2017, our group has been conducting a study that follows up people with hypertension and DM in Northeastern Minas Gerais, Brazil, a resource-constrained region called Vale do Mucuri (Mucuri Valley). Until October 2018, the HealthRise project, funded by the Medtronic Foundation, aimed to improve the screening and disease control of people with hypertension and DM [12]. The main activities were (1) training the multidisciplinary family health team, (2) organizing the flow of spontaneous and scheduled consultations, (3) expanding rational access to complementary exams, (4) supporting group activities, (5) sending text messages to patients' cell phones, and (6) developing and implementing a clinical decision support system (CDSS). Nurses and physicians applied recommendations from evidence-based guidelines in their work routine providing the patient with up-to-date treatment. Community health workers (CHWs) received tablets to enroll patients in the screening phase. These devices were also useful to improve CHWs' work routines, facilitating the entry of data into the Ministry of Health's information system [13].

There was a 2-month transition between the end of the HealthRise project and the beginning of the next project, the Charming Project (Control of Hypertension and Diabetes in Minas Gerais). This transition lasted until December 2018, and, in early 2019, the intervention restarted as the Charming Project, maintaining all the previous components and activities in the same territory.

In the pandemic scenario, which puts at risk the monitoring of those patients in the primary care setting, the purpose of this study was to assess the impact of the COVID-19 pandemic on the frequency of consultations for patients with hypertension and DM and the control of both diseases in a vulnerable region. Additionally, to mitigate the negative impact that the pandemic may have had on these patients, this study evaluated the implementation of a simple and efficient mHealth strategy for CHWs that was used during home visits, to prioritize the in-person consultation of patients with uncontrolled disease.

Methods

Study Design

This mixed methods study was a substudy of the Charming Project that took place in 34 primary health care centers (PHCCs) in 10 municipalities of Mucuri Valley: Ataléia, Catuji, Crisólita, Frei Gaspar, Itaipé, Ladainha, Novo Oriente de Minas, Ouro Verde de Minas, Setubinha, and Teófilo Otoni [13]. Mucuri Valley is part of the Northeast Macro-Region of Minas Gerais, Brazil, with a territorial extension of 24,781.5 km^2 and a population of 516,073 inhabitants, marked by drastic socioeconomic contrasts, high rates of illiteracy and poverty, and low rates of control of hypertension and DM [14].

This study was performed in 4 steps, according to the Medical Research Council framework (Figure 1): (1) identification of gaps in usual care, (2) identification of the components of the intervention through discussions with experts, (3) software development and validation, and (4) pilot testing [15].

Figure 1. Flowchart of the study methodology. DSS: decision support system. PHCC: primary health care center.
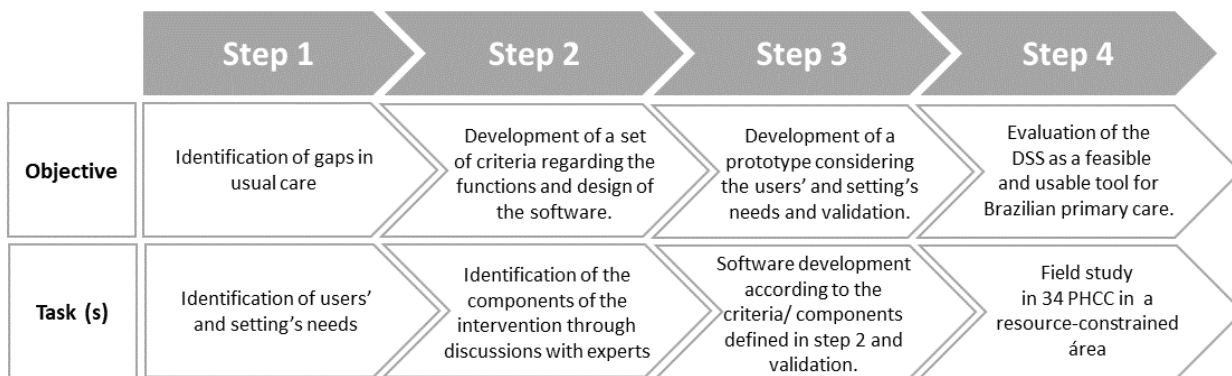
Step 1: Identification of Gaps in Usual Care

Local health authorities confirmed the first case of COVID-19 in Mucuri Valley on April 22, 2020. However, as soon as the World Health Organization (WHO) declared the COVID-19 a global pandemic on March 11, 2020, there was a significant drop in the number of consultations at the PHCC. At the beginning of the pandemic, the Ministry of Health's recommendations were contradictory, and the role of CHWs was not established [16]. In May, general guidelines from the Ministry of Health suspended group activities and home visits, and PHCCs were designated to receive people with respiratory symptoms. Patients with chronic diseases were left in the background, receiving only prescription renewal. Consequently, there was a mischaracterizing of the work routine in the primary care setting.

Quasiexperimental Study

The quasiexperimental study aimed to assess the impact of the COVID-19 pandemic on the frequency of consultations. Likewise, it also aimed to assess the impact of the pandemic on the control of those patients. Therefore, 2 periods were considered: the so-called period 1, from baseline (June 1, 2017) to March 13, 2020, and period 2, from March 14, 2020 (12th epidemiological week, when social restrictions were intensified) to December 31, 2020. This assessment included all patients followed by the HealthRise and Charming Projects with hypertension and DM aged 30 years to 69 years at the project's baseline [12], being monitored in the 34 PHCCs of 10 municipalities who had at least two consultations. The age range was previously determined by the funder and described in the public call for project submission.

Data were obtained through the usual medical and nursing consultation procedures, recorded in the software developed in the study [13]. Variables of interest were sociodemographic data (age, sex, education, income), clinical data (hypertension, DM, stroke, peripheral arterial disease, coronary artery disease, heart failure, alcoholism, physical inactivity, smoking), laboratory data (glycated hemoglobin [HbA_{1c}]), physical examination measures (systolic BP [SBP] and diastolic BP [DBP]), and follow-up data (number of consultations performed). HbA_{1c} was assessed using laboratory tests of peripheral blood samples and point-of-care tests.

Step 2: Identification of the Components of the Intervention

mHealth Solution for CHWs

Since primary health care professionals had already used digital solutions in the previous projects, an mHealth solution was planned for CHW that could identify patients with uncontrolled hypertension and DM at home. From the recognition of these patients, the CHW could prioritize them for medical consultation at the PHCC. The app was developed to run on tablets and smartphones and has a DSS that indicates to the CHW whether the patient's disease is controlled. By entering simple data, the CHW receives immediate feedback during the home visit, providing prompt patient guidance.

The use of synchronous teleconsultations as part of this intervention was considered. However, we detected many barriers that impaired their implementation, such as a significant proportion of illiteracy among the population, social and economic vulnerability, poor internet connectivity in remote areas, and lack of infrastructure.

Procedures

To assess patients with hypertension, the CHW used an automatic arm BP monitor (OMRON HEM-7320). However, in Brazil, CHWs' basic training for their role does not include performing nursing procedures, such as measuring capillary blood glucose [17]. So, in cases of DM, they guided patients to use dipsticks to assess glycosuria in an isolated urine sample. It made screening the most uncontrolled cases (urine glucose concentrations of ++ or more corresponds to blood glucose above 250 mg/dL) possible. In addition, some patients with uncontrolled DM received test strips and instructions to perform glucose self-monitoring and improve the adjustments to the insulin prescription. Using tablets, CHW entered all the information obtained into the app, which consists of a simple questionnaire and a DSS. When the DSS classifies the patient as uncontrolled, a message advises the CHW to make a medical appointment at the PHCC.

Evaluation of glycemic control by HbA_{1c}, using a point-of-care portable HbA_{1c} analyzer, was available in many PHCCs for testing just before the medical consultation, with immediate results, and allowing prompt decision-making.

Prioritization Criteria

Since it would not be possible to attend to all patients with hypertension and DM, prioritization criteria were established according to the findings expected in home visits. The following patients were prioritized for in-person medical consultations at the PHCC: people with hypertension and SBP ≥160 mm Hg or DBP ≥100 mm Hg and/or people with DM and any capillary blood glucose measurement ≥250 mg/dL or glycosuria result ++ or higher.

The intervention was centralized in the CHW because of their close contact with the community through home visits, activities in support groups, and performance of preventive actions. They know about the families' health and social conditions, adherence to treatment, and attendance at consultations at the PHCC [17,18].

Step 3: Software Development and Validation

App Overview

The purpose of the app was to identify, in the home environment, patients who were seriously decompensated or at higher risk of decompensation and to prioritize them for medical consultation at the PHCCs.

To ensure privacy, data are kept encrypted on the tablet. User sessions expire whenever devices enter the sleep mode or periodically (the shortest of times). User digital authentication is secure. The app has features to operate in online and offline modes. Due to the lack of internet connection in patients' homes where data are collected and lack of 3G or 4G connection in the tablets, it is necessary to store patient data used to promptly generate decision support in the device. This information is uploaded online and as soon as CHWs go back to the PHCC.

To develop the app, the developer team and stakeholders had a round of meetings in order to define its scope. Bearing in mind the prospective user profile, Android 4.4 or above (Api level 19, KitKat) was selected as the operating system. A prototype was made and submitted for approval, upon which the development phase began. Netbeans and Java for Android were used together with the libraries Firebase Crashlytics, Analytics, Volley Plus, and Realm Database, the latter for the local mobile

databank. REST was used for mobile-databank (POSTGRESQL) communication. The development phase was incremental, and each successive version was submitted to stakeholders for testing and approval. The last submission included a test battery with final users, the results of which were reported to the developer team for error solution. Once a final submission was approved, the app was released for production. The app can be downloaded from Google Play Store. Support for installation and use was provided by the local team.

The app is freely available in Google Play Store and can be downloaded under the name *Questionário Charming*. At the moment, only the Portuguese version is available.

App Content and Functionality

The app consists of (1) a log-in screen, (2) a patient search screen, (3) a patient registration screen, (4) patient assessment, and (5) decision support.

The login screen allows individualized access to the system through the credentials (user and password) provided to each professional. Once logged in, the professional has access to the screen to search for registered patients (Figure 2).

If the patient was not registered before, the CHW can create a new registration and input demographic data, address, telephone number, and information on diagnosis of hypertension and DM.

After choosing a specific patient, the professional enters the questionnaire screen, which includes BP levels, recent capillary blood glucose levels, glycosuria result, questions about adherence to drug treatment, reasons for nonadherence (if it occurs), if the patient has an insulin prescription, and glucose strip supplies (Figure 3). It is worth mentioning that the Brazilian public health system (Sistema Único de Saúde [SUS]) provides glucometer and blood glucose strips for people with DM who use insulin.

The DSS provides personalized recommendations, generated according to the data entered in the CHW evaluation. The messages alert the CHW if glucose or BP levels are high and suggests scheduling medical or nursing visits in the PHCCs, delivery of drugs or supplies, and prescription renewal (Figure 4).

Figure 2. Example patient search screen with fictitious information: patient's name, registration number in the public health system, birth date, record id, and priority, according to criteria explained in the text.

Figure 3. Patient assessment screen, on which the community health worker easily inputs information obtained during the home visit: blood pressure levels, glycosuria, medication adherence and access, insulin use, and presence of a glucometer at home.

XSL·FO
RenderX

Figure 4. Decision support screen, which shows the best recommendation for the patient after the community health worker inputs and saves the patient data.

Pretesting

In order to ensure that the system was operating as intended, with no bugs, and that the recommendation results matched the prespecified decision tree, the prototype was tested multiple times through manual insertion of test cases. Medical students, professors, and researchers took the tests several times.

Expert Panel Assessment

The expert panel consisted of 4 primary care physicians, 1 nurse, 1 pharmacist, and 1 CHW, all working in primary health care and 2 of them (1 of the physicians and the nurse) also working as health care managers. They were all recognized as technical references and completely independent of the researchers and implementation sites. The specialists tested the app for 1 week by simulating the most different situations that the CHW might

encounter during a home visit. At the end of the tests, they gave their opinion through a questionnaire previously developed by our group [19]. The first part of the questionnaire included the following participant characteristics: sex at birth, age, education level, time since graduation, profession, role in primary care, prior knowledge of information technology (IT), ease of use, and frequency of internet use, for how long during the day, and for what reasons (personal, professional, or other). The second part included Likert scale questions, varying from 1 (strongly disagree) to 5 (strongly agree), to assess feasibility, usability, and utility.

Step 4: Pilot Testing

The IT team and researchers carried out the implementation, through remote and in-person training. CHWs in the 10 municipalities received specific remote training about the (1) correct use of personal protective equipment, (2) identification of suspected cases of COVID-19 during home visits, (3) BP measurement using an automatic arm BP monitor, (4) glycosuria assessment in an isolated urine sample using reagent dipsticks, and (5) identification of cases with uncontrolled hypertension and DM during home visits. Depending on the clinical situation, they had to measure BP and/or glycosuria and guide patients with DM about how to measure capillary blood glucose at home. They underwent a specific training session through a web conference to clarify the expansion of restrictive measures regarding the COVID-19 pandemic. In addition, CHWs received supporting material about the software and instructions for use. The IT team installed the app on the tablets that CHWs were already using. During these visits to the PHCCs, they provided hands-on training on how to use the app and answered CHWs' remaining doubts. Later when devices, app incompatibilities, and login-related issues arose, the IT team promptly solved them remotely.

Statistical Analysis

Continuous variables were described by measures of central tendency (mean or median), dispersion (SD or IQR), and amplitude, according to the distribution assessed using the Kolmogorov-Sminov test. Categorical variables were described with measures of absolute and relative frequencies. Using the Chow test, weekly consultations' time series data were analyzed to identify possible structural changes (ie, sudden changes in the trend of the time series) related to the pandemic. Variables were compared between periods 1 and 2 using Student t tests, Mann-Whitney U tests, or Fisher exact tests, according to the

normality of the distribution and type of variable. Statistical analysis was performed with R software (version 4.0.2) with the strucchange and ggplot2 packages.

Ethical Review

The Federal University of Jequitinhonha and Mucuri Valleys' Research Ethics Committee gave ethical approval for the study, which is registered under the Certificate of Presentation of Ethical Appreciation (CAAE) number 40479820.2.0000.5108.

Results

Quasiexperimental Study

During the 183-week follow-up period ranging from June 2017 to December 2020, 17,345 consultations were carried out, with a median number of 2 consultations per patient, except for those with diagnoses of both hypertension and DM, who had a median of 3 consultations. Physicians (10,199/17,345, 58.8%) performed most of them. HbA_{1c} was evaluated through laboratory tests from peripheral blood samples and point-of-care tests. From the 3488 HbA_{1c} assessments, 1027 (29.4%) used point-of-care tests. Between 2019 and 2020, there was a 74% reduction in the number of HbA_{1c} tests (977 vs 255) and a 62.5% reduction in the number of BP measurements (3898 vs 1461). The overall number of patients was 5202. There were 4936 (94.9%) with hypertension and 1403 (27.0%) with DM. It is noteworthy that DM and hypertension coexisted in 1137 patients (Table 1).

The beginning of the 12th epidemiological week, on March 14, 2020, is a milestone for this study, as it corresponds with the moment of escalation of social restriction measures. The Chow test confirmed the break point, which divided the timeline into period 1 and period 2.

For the before-after assessment, 5070 patients were analyzed. Of these, 4810 (94.9%) patients had hypertension, and 1371 (27.0%) had DM. Among them, 1111 (23.1%) patients had both diseases. Most patients were female (3369/5070, 66.4%), with median age of 56.0 (IQR 48.0-62.0) years and median BMI of 27.9 (IQR 24.6-31.6) kg/m^2 (Table 2).

It is noteworthy that DM and hypertension coexisted in 1111 patients.

There was a significant reduction in the number of consultations, BP measurements, and HbA_{1c} dosage between period 1 and period 2 (Table 3).

Table 1. Consultations, procedures, and patient data.

Characteristic	Overall sample (n=17,345), n (%)	By year, n			
		2017	2018	2019	2020
Number of consultations	17,345 (100)	1990	6325	6437	2593
Number of consultations by health care professional					
Physician	10,201 (58.8)	1300	4393	3450	1058
Nurse	7144 (41.2)	690	1932	2987	1535
Total number of procedures, n (%)	14,584 (84.1)	589	730	1389	780
Procedures, n (%)					
HbA$_{1c}$[a] tests	2461 (16.9)	583	646	977	255
HbA$_{1c}$ POC[b] tests	1027 (7.0)	6	84	412	525
Blood pressure measurements	11,096 (76.1)	1580	4157	3898	1461
Total number of patients, n (%)	5202 (30.0)	1757	3444	2833	1576
Number of patients with each disease, n (%)					
DM[c]	1403 (27.0)	491	911	913	599
Hypertension	4936 (94.9)	1654	3288	2674	1479

[a]HbA$_{1c}$: glycated hemoglobin.

[b]POC: point-of-care.

[c]DM: diabetes.

Table 2. Characteristics at baseline.

Characteristics	Overall (n=5070)		DM[a] (n=1371)		Hypertension (n=4810)	
Gender (female), n (%)	3369 (66.4)		953 (69.5)		3203 (66.6)	
Age (years), median (IQR)	56.0 (48.0-62.0)		57.0 (49.0-63.0)		56.0 (48.0-63.0)	
BMI (kg/m^2), median (IQR)	27.9 (24.6-31.6)[b]		28.5 (25.1-32.4)[c]		27.9 (24.6-31.6)[d]	
Number of consultations, median (IQR)	2.0 (1.0-5.0)		3.0 (2.0-6.0)		2.0 (1.0-5.0)	
HBA$_{1c}$[e] tests						
Number of HbA$_{1c}$ tests, median (IQR)	_[f]	_[f]	1.0 (1.0-3.0)		N/A[g]	N/A
HbA$_{1c}$ result (%), median (IQR)	-	_[f]	7.6 (6.4-9.6)[h]		N/A	N/A
Number of HbA$_{1c}$ results <7%, n (%)	_[f]	_[f]	409 (37.9)[h]		N/A	N/A
BP[i] measures						
Number of BP measures, median (IQR)	_[j]	_[j]	N/A	N/A	1.0 (1.0-3.0)	
SBP[k] (mm Hg), median (IQR)	_[j]	_[j]	N/A	N/A	130.0 (120.0-140.0)[l]	
DBP[m] (mm Hg), median (IQR)	_[j]	_[j]	N/A	N/A	80.0 (80.0-90.0)[l]	
SBP <140 mm Hg and DBP <90 mm Hg, n (%)	_[j]	_[j]	N/A	N/A	1937 (48.7)[l]	

[a]DM: diabetes mellitus.

[b]n=3961 (78.1%).

[c]n=1074 (78.3%).

[d]n=3759 (78.2%).

[e]HbA$_{1c}$: glycated hemoglobin.

[f]Only tested in those with DM, so the values would be the same as those reported under DM.

[g]N/A: not applicable.

[h]n=1079 (78.7%).

[i]BP: blood pressure.

[j]Only tested in those with hypertension, so the values would be the same as those reported under hypertension.

[k]SBP: systolic blood pressure.

[l]n=3980 (82.7%).

[m]DBP: diastolic blood pressure.

Table 3. Weekly number of consultations, blood pressure measurements, and glycated hemoglobin (HbA1c) tests in period 1 and period 2.

Characteristic	Overall (n=183), median (IQR)	Period 1 (n=142) , median (IQR)	Period 2 (n=41) , median (IQR)	P value
Consultations	83.0 (29.0-139.5)	107.0 (60.0-153.0)	20.0 (7.0-29.0)	<.001
BP[a] measures	56.0 (22.5-89.5)	68.0 (38.2-99.0)	11.0 (4.0-19.0)	<.001
HbA$_{1c}$ tests	13.0 (3.5-29.0)	16.5 (5.0-32.8)	4.0 (0.0-22.0)	<.001

[a]BP: blood pressure.

Over the 183 weeks of follow-up in the series, there were other variations in the number of consultations in specific periods. In 2017, with the beginning of patient follow-up, there was a progressive growth in the number of weekly appointments at PHCCs. At the end of the year, there was an abrupt drop in the number of consultations, which is systematically observed in all years. In 2018, the number of consultations tended to remain stable, but there was a sudden drop in October, which corresponded with the transition between the HealthRise and Charming projects, followed by the expected reduction in attendance at the end of the year. In early 2019, when funding was reinstated, the number of consultations rose again. In 2019,

the pattern of consultations presented the usual variations, and in the first 2 months of 2020, activities in the PHCCs were normal. However, at the beginning of the 12th epidemiological week, there was a dramatic reduction in the number of consultations at PHCCs. On April 22, 2020, the first COVID-19 case in the region was confirmed. Throughout 2020, a reduced number of weekly consultations is clearly observed (Figure 5).

Of the 5070 patients who were being followed since before the pandemic, 4298 (84.8%) did not return for a consultation after the social distancing measures were implemented. Of the 772 (15.2%) individuals who returned, the median time between the

last consultation before the social distancing measures and the first consultation after it was 233 days (Table 4).

It is noteworthy that DM and hypertension coexisted in 313 patients who returned for a consultation in period 2.

The proportion of patients who returned after period 2 was much lower than the proportion observed during period 1 (772/5070, 15.2% vs 2565/5070, 50.6%; Table 5).

It is noteworthy that DM and hypertension coexisted in 1111 patients.

The characteristics common to the groups of those who did not return for consultation in period 2 and those who did return were compared. The median age of those who returned was slightly higher—58.0 (IQR 51.0-65.0) years versus 56.0 (IQR 48.0-63.0) years (*P*<.001)—but there was no difference regarding their sex. The median number of consultations in

period 1 was also higher in that population: median 5.0 (IQR 3.0-7.0) vs 2.0 (IQR 1.0-4.0). Lower median SBP levels—120.0 (IQR 120.0-140.0) mm Hg versus 130.0 (IQR 120.0-140.0) mm Hg—and median DBP levels—80.0 (IQR 80.0-80.0) mm Hg versus 80.0 (IQR 80.0-90.0) mm Hg—were found in patients who returned for consultation, as well as a higher proportion of patients with controlled hypertension (431/772 64.3% vs 1735/4298, 52.8%). However, among people with DM, there was no difference in HbA_{1c} levels between periods 1 and 2 (Table 6).

It is noteworthy that DM and hypertension coexisted in 798 patients who did not return and 313 patients who returned.

Since June 2017, 110 patients were discharged from the system: 35 due to formal withdrawal, 62 who moved, 12 who died from a nontraumatic cause, and 1 from a traumatic cause.

Figure 5. Weekly number of consultations and variations between 2017 and 2020.
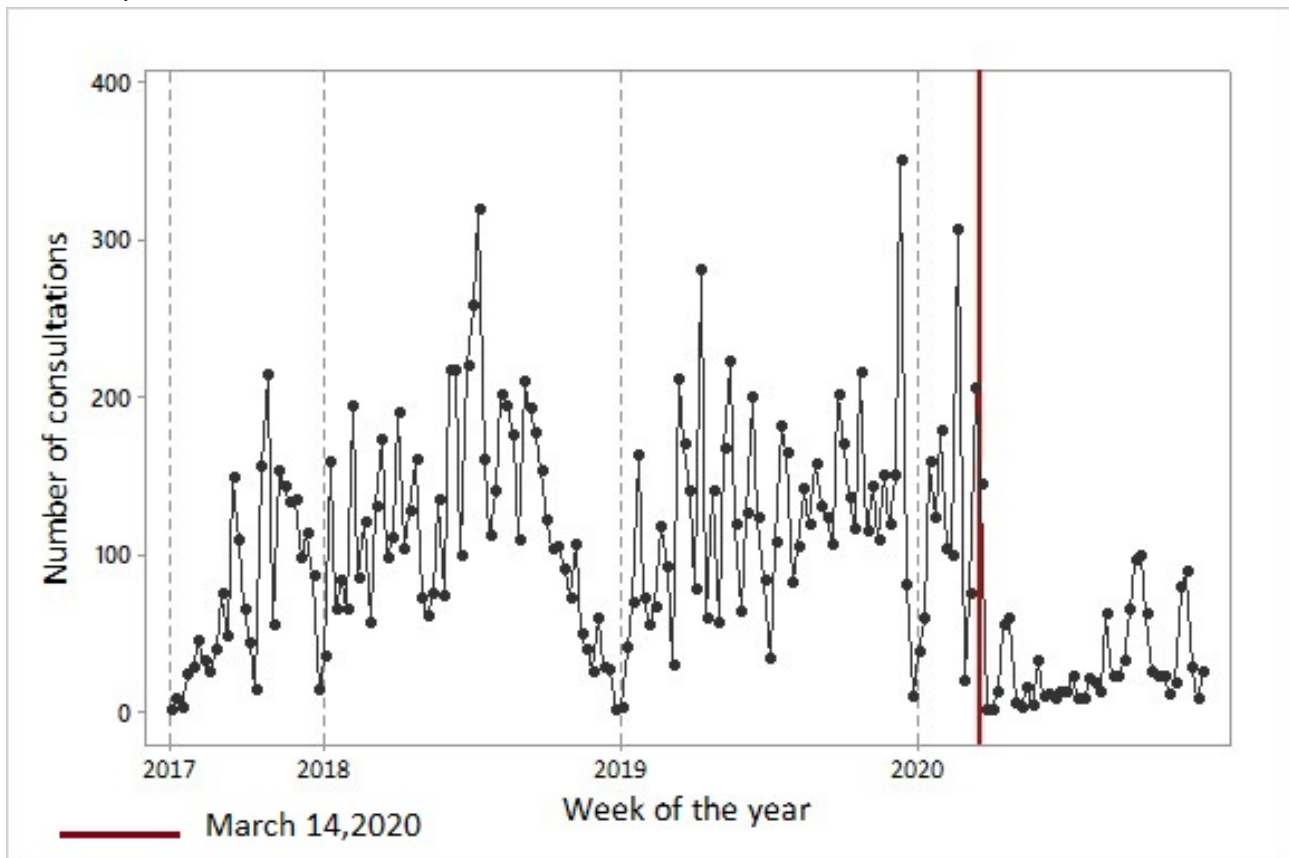
Table 4. Time between the last consultation in period 1 and the first consultation in period 2.

Measurement	Overall (n=772)	DM[a] (n=366)	Hypertension (n=719)
Time (days), median (IQR)	233.0 (122.8-356.0)	206.0 (114.0-306.5)	238.0 (125.0-362.0)

[a]DM: diabetes mellitus.

Table 5. Comparison of returns for consultation at a primary health care center (PHCC) in periods 1 and 2.

Characteristic	Overall (n=5070)	DM[a] (n=1371)	Hypertension (n=4810)
No return, n (%)	1732 (34.2)	297 (21.7)	1660 (34.5)
Return in period 1, n (%)	2566 (50.6)	708 (51.6)	2431 (50.5)
Return in period 2, n (%)	772 (15.2)	366 (26.7)	719 (14.9)

[a]DM: diabetes mellitus.

Table 6. Comparison of the characteristics of patients between those who returned and did not return in period 2.

Characteristic	Did not return (n=4298)	Returned (n=772)	*P* value
Gender, female n (%)	2849 (66.3)	520 (67.4)	.56
Age (years), median (IQR)	56.0 (48.0-63.0)	58.0 (51.0-65.0)	<.001
Disease, n (%)			
Hypertension	4091 (95.2)	719 (93.1)	<.001
DM[a]	1005 (23.4)	366 (47.4)	
BMI[b], median (IQR)	27.9 (24.7-31.6)[c]	27.8 (24.6- 31.8)[d]	.88
Number of consultations before the pandemic, median (IQR)	2.0 (1.0-4.0)	5.0 (3.0-7.0)	<.001
BP[e] values			
SBP[f] (mm Hg)[b], median (IQR)	130.0 (120.0-140.0)[g]	120.0 (120.0-140.0)[h]	<.001
DBP[i] (mm Hg)[b], median (IQR)	80.0 (80.0-90.0)[g]	80.0 (80.0-80.0)[h]	<.001
SBP <140 mm Hg and DBP <90 mm Hg, n (%)[b]	1735 (52.8)[g]	431 (64.3)[h]	<.001
HbA$_{1c}$[j]			
HbA$_{1c}$ result (%), median (IQR)[b]	7.6 (6.3-9.5)[k]	7.6 (6.5-9.3)[l]	.80
HbA$_{1c}$ <7%, n (%)[b]	272 (37.3)[k]	108 (33.8)[l]	.28

[a]DM: diabetes mellitus.

[b]Last measure or test before the pandemic.

[c]n=3243 (75.2%).

[d]n=685 (88.7%).

[e]BP: blood pressure.

[f]SBP: systolic blood pressure.

[g]n=3286 (76.5%).

[h]n=670 (86.8%).

[i]DBP: diastolic blood pressure.

[j]HbA$_{1c}$: glycated hemoglobin.

[k]n=730 (73%).

[l]n=320 (87%).

Validation: Expert Panel Assessment

All 7 experts who participated fully agreed that the app could be used in primary care settings to improve care for people with hypertension and/or DM. They also agreed that it could be easily incorporated in work routines (median 4.0, IQR 4.0-5.0). All believed that the app does not cause significant delays in the daily routine. As for usability, the overall evaluation was good, but all professionals claimed that the app was not intuitive and previous training is necessary. As for utility, they believed that the app might improve the treatment and care of people with hypertension and DM. The characteristics of the experts and the results of the feasibility assessment can be seen in Multimedia Appendix 1 and Multimedia Appendix 2.

Pilot Testing

From November 2020 to May 2021, 211 CHWs from 10 municipalities received training to evaluate patients with hypertension and DM during in-home visits. The implementation was carried out progressively, and the first records were on paper. With hands-on training in May 2021, CHWs started using the app and entering data obtained from home visits. From May 2021 to December 2021, there were 1314 records from CHWs' in-home visits using the app: 1266 (96.4%) were patients with

hypertension, 245 (18.6%) were patients with DM, and 197 (15.0%) were patients with both diseases. They found 220 (220/1266, 17.38%) patients with high BP (≥160 mm Hg or ≥100 mm Hg); 34 (34/245, 13.9%) patients reported capillary blood glucose level at any time in the last 20 days ≥250 mg/dL. Regarding glycosuria, 40 (40/245, 16.3%) had a positive result (Multimedia Appendix 3).

Discussion

Main Findings

We observed a significant reduction in the weekly number of consultations, BP measurements, and HbA_{1c} tests between periods 1 and 2. There was a systematic reduction in PHCC visits in all the last months of the years, regardless of the pandemic, due to the December holidays and summer vacations. The proportion of patients who returned for a medical consultation at the PHCCs was also significantly lower during the pandemic. Of those who returned, it was observed that they were more assiduous in follow-up appointments before the pandemic when compared to the period after isolation measures were implemented. Lower SBP and DBP levels were observed in patients who returned for consultation during the pandemic, as well as a higher proportion of patients with controlled hypertension. However, among people with DM, there was no difference in HbA_{1c} levels between periods 1 and 2.

The reduction in the number of consultations was observed globally. According to a survey by the WHO, completed by 155 countries in May 2020, 53% of them had partially or completely disrupted services for hypertension treatment and 49% for DM and DM-related complications at that time [20]. Another survey that included 202 health care professionals from 47 countries observed that DM was the chronic disease most affected by the reduction in health care resources due to COVID-19 [21]. A worldwide survey submitted from 909 centers performing cardiac diagnostic procedures in 108 countries found a 42% drop in the number of procedures from March 2019 to March 2020 and a 64% drop from March 2019 to April 2020 [22].

In Germany, during the lockdown, there was a dramatic reduction in the number of consultations with general practitioners, independent of age, sex, and location (rural vs urban areas) [23]. The immediate need for social restriction forced health care systems to adopt telemedicine for ensuring baseline and, in some selected cases, advanced health care support [24]. As a result, the COVID-19 pandemic has strengthened the use of telemedicine as an indispensable resource to monitor the health conditions of people at home, including those with hypertension [25]. In Italy, the number of visits to general practitioners' offices and tests dropped markedly. At the same time, the number of home app users, exchanging data between patients and doctors, significantly increased. It thus resulted in significant improvement of BP control [26]. Telemedicine with video consultations was a significantly effective tool in the management of people with hypertension, with high levels of patient satisfaction in the United States [27]. Telemedicine consultations contributed to maintaining care for 69% people with antineutrophil cytoplasmic antibody–associated vasculitis, both in the United States and the United Kingdom, allowing monitoring and identification of deterioration at home [28]. With regards to DM, the role of telehealth in the care of people with type 1 DM has expanded dramatically during the COVID-19 pandemic [29]. Successful experiences have been described in India and Australia, where patients who attended through teleconsultations had slightly better glycemic control than those in the pre-COVID-19 period [30,31]. In the United States, although DM-related outpatient visits and testing fell during the pandemic, there was no evidence of a negative association with glycemic control. Telemedicine may have prevented substantive disruptions in medication prescribing [32].

In Brazil, the Ministry of Health has taken steps to ensure care for people with respiratory syndrome throughout the SUS health network, reinforcing primary care as the preferred gateway [33]. However, at the same time, the heterogeneity of the organization of primary care in Brazil [34] as well as the restrictions imposed by the pandemic [35] compromise the continuous monitoring of patients with chronic diseases. CHWs were prioritized for health surveillance and administrative actions within the PHCCs and even issues that are not their responsibility, such as helping with vaccination campaigns [16]. The problem was even more serious in remote and resource-constrained areas, where patients need to travel long distances to receive health care and medication. As the management of COVID-19 cases has become a priority for most health units, nonemergency medical services or services not related to COVID-19 have been postponed indefinitely, predisposing patients with hypertension and DM to a high potential for increased risk of complications and a worse prognosis [35].

The consequences of these changes can already be seen in Brazil. In the 6 capital cities that had the highest number of deaths from COVID-19, there was also an increase in the number of deaths from cardiovascular diseases. These findings were probably a consequence of the poor health services infrastructure and reflects the increase in home deaths due to the impaired access to health services [36]. Cardiovascular diseases have a significant relationship with COVID-19, both as risk factors and prognostic indicators and as complications [37,38]. Nevertheless, despite this, the pandemic forced a reorganization of health systems that compromised the provision of adequate and timely health services [39,40] and negatively affected mortality [41].

Facing the reality of patients at home and without proper care, a strategy using mHealth for CHWs was planned and implemented to identify patients at risk. These patients were referred to the PHCCs for medical consultation. CHWs were chosen as the central point of this intervention because they represent the bridge between the health system and the users, especially the most vulnerable, enabling the capillarity of the SUS. They live in the same community for which they care and ensure that the family health strategy is taken to the communities. Their responsibilities include education and health promotion, keeping records of individuals and families, identifying those at risk, making regular home visits to monitor children's vaccinations or the well-being of chronic patients, scheduling consultations with a maternal health specialist,

advising on the correct use of medications, and contributing to mosquito control campaigns [42].

Although digital health was in use long before the pandemic, the challenges of the current moment encouraged its escalation, which is happening quickly. The options for digital tools are vast, including video visits, email, mobile phone apps, chatbots, voice interface systems, smartwatches, oxygen monitors, and thermometers [11]. mHealth technology has been proven to enhance the management of patients with chronic diseases [43] and is associated with a reduction in BP and better medication adherence for people with hypertension [44]. Also, mHealth‐based interventions for patients with type 1 DM significantly decreased HbA_{1c}, improved life satisfaction, and improved mental health [45]

The use of digital tools by CHWs has been successfully carried out in several countries, not only to fight the pandemic but also to minimize the impact of interruptions in patient follow-up [46]. However, to be effective, digital health solutions for CHWs need to be easy to use. Furthermore, there is a need to improve their skills, improve accessibility, and ensure continuity of care [47]. For this, challenges such as training on new mHealth solutions, weak technical support, and issues of internet connectivity must be overcome [10,48]. The app is an easy-to-use digital tool, with objective questions and yes/no/don't know answers. Therefore, education level is not an obstacle to using it. Likewise, we guarantee training and support for the use of the app, solving doubts online in real time.

The expert panel assessment classified the app as feasible and useful. However, they agreed that its use requires training. They believe in the app's potential to contribute to the identification of patients at risk and the provision of medication by health managers. This information is important to obtain regardless of the pandemic. The federal government guarantees the donation of glucometers for people with DM who use insulin, but the supply of strips is erratic. The app can provide important feedback regarding municipal financing and logistics for purchasing medicines. The measurement of BP by the CHW using a digital monitor adds value to the home visit. The professionals and the patients have accepted this innovation without restrictions.

The pilot test results reflect a short evaluation period as the intervention itself is ongoing. The findings showed a low percentage of patients with uncontrolled hypertension and DM when evaluated at home. Medication adherence as well as the supply of medications were satisfactory. The aim of measuring BP, glycosuria, and blood glucose levels was to prioritize in-person medical consultation for those at risk and not just to identify people with uncontrolled disease. The major limitation was that the PHCCs were designed to receive people with respiratory symptoms as a priority. Therefore, it was not possible

to schedule a medical consultation for all patients with hypertension and DM outside the control goals. As the app was designed to be simple, the CHW does not enter the patient's BP but only marks if it is ≥160 x 100 mm Hg. Therefore, people outside the control target whose BP is <160 x 100 mm Hg will not be referred to a PHCC, which can underestimate the number of patients outside the control goals. The same is true for DM, as only those with a change equivalent to blood glucose ≥250 mg/dL will be identified and referred for consultation. Another aspect that needs to be considered is that patients who participated in the pilot test are the ones who most frequently attended the PHCC. The researchers did not interfere in the choice of patients for the pilot test, but the CHW preferred to visit those with greater adherence to treatment. This justifies the finding of a low proportion of people with BP and glucose levels high enough to warrant a medical consultation and justifies the adequate supplies of medication.

Limitations

The cutoff BP, glycosuria, and blood glucose values to indicate the need for a medical consultation were established to prioritize those most in need, since the PHCC could not attend to all patients. Therefore, the results cannot be seen as just a definition of the proportion of patients with uncontrolled disease, as this underestimates the real number of patients with uncontrolled disease.

Next Steps

The integration of the CDSS with the current electronic medical record of the Brazilian public health system software (e-SUS) is a challenge, and we are currently working to overcome this barrier. We are planning to expand the project to PHCCs in other municipalities, even with the reduction of the pandemic. Our team believes that this is a promising strategy that values and expands the CHWs' skills, strengthens their relationship with the community, promotes the identification and adequate referral of patients with uncontrolled disease, and brings the users even closer to the SUS. We expect services to be fully resumed in the near future, when we will be able to review the cut-off levels of BP and glycosuria or glycemia to refer to the PHCC.

Conclusions

The COVID-19 pandemic caused significant impairment in the follow-up of patients with hypertension and DM in a resource-constrained region of Brazil, due to the reduction in the number of medical and nursing consultations, BP measures, and HbA_{1c} assessments after the initiation of social restriction measures. A DSS app has proven to be feasible and useful, and CHWs are using it to identify patients with uncontrolled hypertension and DM who are at home and at risk.

XSL•FO

RenderX

Authors' Contributions

All authors made substantial contributions to the conception or design of the work and critically revised the manuscript for important intellectual content. CCRC, MCP, and MSM made substantial contributions to the acquisition, analysis, or interpretation of data for the work. CCRC, MSM, and ALPR provided final approval of the version to be published. CCRM and MSM agree 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.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Characteristics of Experts who validated the app (N=7).
[DOCX File , 16 KB - medinform_v10i3e35216_app1.docx]

Multimedia Appendix 2
Feasibility, usability and utility assessment.
[DOCX File , 18 KB - medinform_v10i3e35216_app2.docx]

Multimedia Appendix 3
Characteristics of the study patients.
[DOCX File , 22 KB - medinform_v10i3e35216_app3.docx]

References

1. Puntillo F, Giglio M, Brienza N, Viswanath O, Urits I, Kaye AD, et al. Impact of COVID-19 pandemic on chronic pain management: Looking for the best way to deliver care. Best Pract Res Clin Anaesthesiol 2020 Sep;34(3):529-537 [FREE Full text] [doi: 10.1016/j.bpa.2020.07.001] [Medline: 33004164]
2. Palmer K, Monaco A, Kivipelto M, Onder G, Maggi S, Michel J, et al. The potential long-term impact of the COVID-19 outbreak on patients with non-communicable diseases in Europe: consequences for healthy ageing. Aging Clin Exp Res 2020 Jul;32(7):1189-1194 [FREE Full text] [doi: 10.1007/s40520-020-01601-4] [Medline: 32458356]
3. Anjana RM, Mohan V, Rangarajan S, Gerstein HC, Venkatesan U, Sheridan P, et al. Contrasting associations between diabetes and cardiovascular mortality rates in low-, middle-, and high-income countries: cohort study data from 143,567 individuals in 21 countries in the PURE study. Diabetes Care 2020 Dec;43(12):3094-3101 [FREE Full text] [doi: 10.2337/dc20-0886] [Medline: 33060076]
4. Barroso WKS, Rodrigues CIS, Bortolotto LA, Mota-Gomes MA, Brandão AA, Feitosa ADDM, et al. Brazilian Guidelines of Hypertension - 2020. Arq Bras Cardiol 2021 Mar;116(3):516-658 [FREE Full text] [doi: 10.36660/abc.20201238] [Medline: 33909761]
5. Pranata R, Lim MA, Huang I, Raharjo SB, Lukito AA. Hypertension is associated with increased mortality and severity of disease in COVID-19 pneumonia: A systematic review, meta-analysis and meta-regression. J Renin Angiotensin Aldosterone Syst 2020 May 14;21(2):1470320320926899 [FREE Full text] [doi: 10.1177/1470320320926899] [Medline: 32408793]
6. Huang I, Lim MA, Pranata R. Diabetes mellitus is associated with increased mortality and severity of disease in COVID-19 pneumonia - A systematic review, meta-analysis, and meta-regression. Diabetes Metab Syndr 2020 Jul;14(4):395-403 [FREE Full text] [doi: 10.1016/j.dsx.2020.04.018] [Medline: 32334395]
7. Sun S, Folarin AA, Ranjan Y, Rashid Z, Conde P, Stewart C, RADAR-CNS Consortium. Using smartphones and wearable devices to monitor behavioral changes during COVID-19. J Med Internet Res 2020 Sep 25;22(9):e19992 [FREE Full text] [doi: 10.2196/19992] [Medline: 32877352]

XSL•FO
RenderX

8.   Bakaloudi DR, Jeyakumar DT, Jayawardena R, Chourdakis M. The impact of COVID-19 lockdown on snacking habits, fast-food and alcohol consumption: A systematic review of the evidence. Clin Nutr 2021 Apr 17:1 [FREE Full text] [doi: 10.1016/j.clnu.2021.04.020] [Medline: 34049747]

9.   Bhutani S, Cooper JA. COVID-19-related home confinement in adults: weight gain risks and opportunities. Obesity (Silver Spring) 2020 Sep;28(9):1576-1577 [FREE Full text] [doi: 10.1002/oby.22904] [Medline: 32428295]

10. Winters N, O'Donovan J, Geniets A. A new era for community health in countries of low and middle income? Lancet Glob Health 2018 May;6(5):e489-e490 [FREE Full text] [doi: 10.1016/S2214-109X(18)30072-X] [Medline: 29653617]

11. Golinelli D, Boetto E, Carullo G, Nuzzolese AG, Landini MP, Fantini MP. Adoption of digital technologies in health care during the COVID-19 pandemic: systematic review of early scientific literature. J Med Internet Res 2020 Nov 06;22(11):e22280 [FREE Full text] [doi: 10.2196/22280] [Medline: 33079693]

12. Flor LS, Wilson S, Bhatt P, Bryant M, Burnett A, Camarda JN, et al. Community-based interventions for detection and management of diabetes and hypertension in underserved communities: a mixed-methods evaluation in Brazil, India, South Africa and the USA. BMJ Glob Health 2020 Jun;5(6):1 [FREE Full text] [doi: 10.1136/bmjgh-2019-001959] [Medline: 32503887]

13. Marcolino MS, Oliveira JAQ, Cimini CCR, Maia JX, Pinto VSOA, Sá TQV, et al. Development and implementation of a decision support system to improve control of hypertension and diabetes in a resource-constrained area in Brazil: mixed methods study. J Med Internet Res 2021 Jan 11;23(1):e18872 [FREE Full text] [doi: 10.2196/18872] [Medline: 33427686]

14. Territorial Areas. Instituto Brasileiro de Geografia e Estatística. URL: https://www.ibge.gov.br/en/geosciences/ territorial-organization/territorial-organization/18092-territorial-areas.html?=&t=downloads [accessed 2022-03-07]

15. Craig P, Dieppe P, Macintyre S, Michie S, Nazareth I, Petticrew M. Developing and evaluating complex interventions: the new Medical Research Council guidance. Int J Nurs Stud 2013 May;50(5):587-592. [doi: 10.1016/j.ijnurstu.2012.09.010] [Medline: 23159157]

16. Lotta G, Wenham C, Nunes J, Pimenta DN. Community health workers reveal COVID-19 disaster in Brazil. The Lancet 2020 Aug;396(10248):365-366. [doi: 10.1016/s0140-6736(20)31521-x] [Medline: 32659212]

17. Krieger MGM, Wenham C, Nacif Pimenta D, Nkya TE, Schall B, Nunes AC, et al. How do community health workers institutionalise: An analysis of Brazil's CHW programme. Glob Public Health 2021 Jun 23:1-18. [doi: 10.1080/17441692.2021.1940236] [Medline: 34161201]

18. Santos ADFD, Rocha HAD, Lima MDLDD, Abreu DMXD, Silva A, de Araújo LHL, et al. Contribution of community health workers to primary health care performance in Brazil. Rev Saude Publica 2020 Dec 12;54:143 [FREE Full text] [doi: 10.11606/s1518-8787.2020054002327] [Medline: 33331421]

19. Silveira DV, Marcolino MS, Machado EL, Ferreira CG, Alkmim MBM, Resende ES, et al. Development and evaluation of a mobile decision support system for hypertension management in the primary care setting in Brazil: mixed-methods field study on usability, feasibility, and utility. JMIR Mhealth Uhealth 2019 Mar 25;7(3):e9869 [FREE Full text] [doi: 10.2196/mhealth.9869] [Medline: 30907740]

20. COVID-19 significantly impacts health services for noncommunicable diseases. World Health Organization. 2020 Jun 01. URL: https://www.who.int/news/item/ 01-06-2020-covid-19-significantly-impacts-health-services-for-noncommunicable-diseases [accessed 2022-03-07]

21. Chudasama YV, Gillies CL, Zaccardi F, Coles B, Davies MJ, Seidu S, et al. Impact of COVID-19 on routine care for chronic diseases: A global survey of views from healthcare professionals. Diabetes Metab Syndr 2020;14(5):965-967 [FREE Full text] [doi: 10.1016/j.dsx.2020.06.042] [Medline: 32604016]

22. Einstein AJ, Shaw LJ, Hirschfeld C, Williams MC, Villines TC, Better N, the, INCAPS COVID Investigators Group. International impact of COVID-19 on the diagnosis of heart disease. J Am Coll Cardiol 2021 Jan 19;77(2):173-185 [FREE Full text] [doi: 10.1016/j.jacc.2020.10.054] [Medline: 33446311]

23. Schäfer I, Hansen H, Menzel A, Eisele M, Tajdar D, Lühmann D, et al. The effect of COVID-19 pandemic and lockdown on consultation numbers, consultation reasons and performed services in primary care: results of a longitudinal observational study. BMC Fam Pract 2021 Jun 23;22(1):125 [FREE Full text] [doi: 10.1186/s12875-021-01471-3] [Medline: 34162343]

24. Citoni B, Figliuzzi I, Presta V, Volpe M, Tocci G. Home blood pressure and telemedicine: a modern approach for managing hypertension during and after COVID-19 pandemic. High Blood Press Cardiovasc Prev 2022 Jan;29(1):1-14 [FREE Full text] [doi: 10.1007/s40292-021-00492-4] [Medline: 34855154]

25. Omboni S, McManus RJ, Bosworth HB, Chappell LC, Green BB, Kario K, et al. Evidence and recommendations on the use of telemedicine for the management of arterial hypertension: an international expert position paper. Hypertension 2020 Nov;76(5):1368-1383. [doi: 10.1161/HYPERTENSIONAHA.120.15873] [Medline: 32921195]

26. Omboni S, Ballatore T, Rizzi F, Tomassini F, Panzeri E, Campolo L. Telehealth at scale can improve chronic disease management in the community during a pandemic: An experience at the time of COVID-19. PLoS One 2021;16(9):e0258015 [FREE Full text] [doi: 10.1371/journal.pone.0258015] [Medline: 34587198]

27. Taylor P, Berg C, Thompson J, Dean K, Yuan T, Nallamshetty S, et al. Effective access to care in a crisis period: hypertension control during the COVID-19 pandemic by telemedicine. Mayo Clin Proc Innov Qual Outcomes 2022 Feb;6(1):19-26 [FREE Full text] [doi: 10.1016/j.mayocpiqo.2021.11.006] [Medline: 34805763]

28. Kant S, Morris A, Ravi S, Floyd L, Gapud E, Antichos B, et al. The impact of COVID-19 pandemic on patients with ANCA associated vasculitis. J Nephrol 2021 Feb;34(1):185-190 [FREE Full text] [doi: 10.1007/s40620-020-00881-3] [Medline: 33034038]

29. Kompala T, Neinstein AB. Telehealth in type 1 diabetes. Curr Opin Endocrinol Diabetes Obes 2021 Feb 01;28(1):21-29. [doi: 10.1097/MED.0000000000000600] [Medline: 33332927]

30. Gopalan HS, Haque I, Ahmad S, Gaur A, Misra A. "Diabetes care at doorsteps": A customised mobile van for the prevention, screening, detection and management of diabetes in the urban underprivileged populations of Delhi. Diabetes Metab Syndr 2019;13(6):3105-3112. [doi: 10.1016/j.dsx.2019.11.008] [Medline: 31790964]

31. Wong VW, Wang A, Manoharan M. Utilisation of telehealth for outpatient diabetes management during COVID-19 pandemic: how did the patients fare? Intern Med J 2021 Dec;51(12):2021-2026 [FREE Full text] [doi: 10.1111/imj.15441] [Medline: 34227718]

32. Patel SY, McCoy RG, Barnett ML, Shah ND, Mehrotra A. Diabetes care and glycemic control during the COVID-19 pandemic in the United States. JAMA Intern Med 2021 Oct 01;181(10):1412-1414. [doi: 10.1001/jamainternmed.2021.3047] [Medline: 34228043]

33. National Contingency Plan for Human Infection by the new Coronavirus (Plano de Contingência Nacional para Infecção Humana pelo novo Coronavírus) COVID-19. Ministério da Saúde. 2020. URL: https://pesquisa.bvsalud.org/portal/resource/en/biblio-1052499 [accessed 2022-03-07]

34. Engstrom EM, Melo EA, Giovanella L, Mendes AMM, Grabois V, de Mendonça MHM. Organização da atenção primária à saúde no SUS no enfrentamento da covid-19. Organização do cuidado na pandemia de covid-19. Rio de Janeiro: Editora Fiocruz; 2020 May 21. URL: https://www.arca.fiocruz.br/handle/icict/41404 [accessed 2021-12-03]

35. The impact of the COVID-19 pandemic on noncommunicable disease resources and services: results of a rapid assessment. World Health Organization. 2020 Sep 03. URL: https://apps.who.int/iris/handle/10665/334136 [accessed 2021-12-06]

36. Brant LCC, Nascimento BR, Teixeira RA, Lopes MACQ, Malta DC, Oliveira GMM, et al. Excess of cardiovascular deaths during the COVID-19 pandemic in Brazilian capital cities. Heart 2020 Dec;106(24):1898-1905 [FREE Full text] [doi: 10.1136/heartjnl-2020-317663] [Medline: 33060261]

37. Golemi Minga I, Golemi L, Tafur A, Pursnani A. The novel coronavirus disease (COVID-19) and its impact on cardiovascular disease. Cardiol Rev 2020;28(4):163-176. [doi: 10.1097/CRD.0000000000000317] [Medline: 32427637]

38. Bonow RO, Fonarow GC, O'Gara PT, Yancy CW. Association of coronavirus disease 2019 (COVID-19) with myocardial injury and mortality. JAMA Cardiol 2020 Jul 01;5(7):751-753. [doi: 10.1001/jamacardio.2020.1105] [Medline: 32219362]

39. Wadhera RK, Shen C, Gondi S, Chen S, Kazi DS, Yeh RW. Cardiovascular deaths during the COVID-19 pandemic in the United States. J Am Coll Cardiol 2021 Jan 19;77(2):159-169 [FREE Full text] [doi: 10.1016/j.jacc.2020.10.055] [Medline: 33446309]

40. Wu J, Mamas MA, Mohamed MO, Kwok CS, Roebuck C, Humberstone B, et al. Place and causes of acute cardiovascular mortality during the COVID-19 pandemic. Heart 2021 Jan 28;107(2):113-119 [FREE Full text] [doi: 10.1136/heartjnl-2020-317912] [Medline: 32988988]

41. Greenberg A, Pemmasani G, Yandrapalli S, Frishman WH. Cardiovascular and cerebrovascular complications with COVID-19. Cardiol Rev 2021;29(3):143-149 [FREE Full text] [doi: 10.1097/CRD.0000000000000385] [Medline: 33758123]

42. Nunes J, Lotta G. Discretion, power and the reproduction of inequality in health policy implementation: Practices, discursive styles and classifications of Brazil's community health workers. Soc Sci Med 2019 Dec;242:112551 [FREE Full text] [doi: 10.1016/j.socscimed.2019.112551] [Medline: 31622914]

43. Smith JC, Schatz BR. Feasibility of mobile phone-based management of chronic illness. AMIA Annu Symp Proc 2010 Nov 13;2010:757-761 [FREE Full text] [Medline: 21347080]

44. Xu H, Long H. The effect of smartphone app-based interventions for patients with hypertension: systematic review and meta-analysis. JMIR Mhealth Uhealth 2020 Oct 19;8(10):e21759 [FREE Full text] [doi: 10.2196/21759] [Medline: 33074161]

45. Chin-Jung L, Hsiao-Yean C, Yeu-Hui C, Kuan-Chia L, Hui-Chuan H. Effects of mobile health interventions on improving glycemic stability and quality of life in patients with type 1 diabetes: A meta-analysis. Res Nurs Health 2021 Feb;44(1):187-200. [doi: 10.1002/nur.22094] [Medline: 33368403]

46. Feroz AS, Khoja A, Saleem S. Equipping community health workers with digital tools for pandemic response in LMICs. Arch Public Health 2021 Jan 04;79(1):1 [FREE Full text] [doi: 10.1186/s13690-020-00513-z] [Medline: 33390163]

47. Maciel FBM, Santos HLPCD, Carneiro RADS, Souza EAD, Prado NMDBL, Teixeira CFDS. Community health workers: reflections on the health work process in Covid-19 pandemic times. Cien Saude Colet 2020 Oct;25(suppl 2):4185-4195 [FREE Full text] [doi: 10.1590/1413-812320202510.2.28102020] [Medline: 33027355]

48. Feroz A, Jabeen R, Saleem S. Using mobile phones to improve community health workers performance in low-and-middle-income countries. BMC Public Health 2020 Jan 13;20(1):49 [FREE Full text] [doi: 10.1186/s12889-020-8173-3] [Medline: 31931773]

Abbreviations

BP: blood pressure

CDSS: clinical decision support system
CHW: community health workers
DBP: diastolic blood pressure
DM: diabetes mellitus
DSS: decision support system
HbA1c: glycated hemoglobin
IT: information technology
mHealth: mobile health
NCD: noncommunicable chronic disease
PHCC: primary health care center
SBP: systolic blood pressure
SUS: Sistema Único de Saúde (Brazilian public health system)
WHO: World Health Organization

XSL·FO
RenderX

<u>Original Paper</u>

Strategies to Address the Lack of Labeled Data for Supervised Machine Learning Training With Electronic Health Records: Case Study for the Extraction of Symptoms From Clinical Notes

Marie Humbert-Droz[1], PhD; Pritam Mukherjee[1], PhD; Olivier Gevaert[1,2], PhD

[1]Stanford Center for Biomedical Informatics Research, Department of Medicine, Stanford University, Stanford, CA, United States
[2]Department of Biomedical Data Science, Stanford University, Stanford, CA, United States

Corresponding Author:
Olivier Gevaert, PhD
Stanford Center for Biomedical Informatics Research
Department of Medicine
Stanford University
Medical School Office Building
Stanford, CA, 94305
United States
Phone: 1 650 721 2378
Email: olivier.gevaert@stanford.edu

Abstract

Background: Automated extraction of symptoms from clinical notes is a challenging task owing to the multidimensional nature of symptom description. The availability of labeled training data is extremely limited owing to the nature of the data containing protected health information. Natural language processing and machine learning to process clinical text for such a task have great potential. However, supervised machine learning requires a great amount of labeled data to train a model, which is at the origin of the main bottleneck in model development.

Objective: The aim of this study is to address the lack of labeled data by proposing 2 alternatives to manual labeling for the generation of training labels for supervised machine learning with English clinical text. We aim to demonstrate that using lower-quality labels for training leads to good classification results.

Methods: We addressed the lack of labels with 2 strategies. The first approach took advantage of the structured part of electronic health records and used diagnosis codes (International Classification of Disease–10th revision) to derive training labels. The second approach used weak supervision and data programming principles to derive training labels. We propose to apply the developed framework to the extraction of symptom information from outpatient visit progress notes of patients with cardiovascular diseases.

Results: We used >500,000 notes for training our classification model with International Classification of Disease–10th revision codes as labels and >800,000 notes for training using labels derived from weak supervision. We show that the dependence between prevalence and recall becomes flat provided a sufficiently large training set is used (>500,000 documents). We further demonstrate that using weak labels for training rather than the electronic health record codes derived from the patient encounter leads to an overall improved recall score (10% improvement, on average). Finally, the external validation of our models shows excellent predictive performance and transferability, with an overall increase of 20% in the recall score.

Conclusions: This work demonstrates the power of using a weak labeling pipeline to annotate and extract symptom mentions in clinical text, with the prospects to facilitate symptom information integration for a downstream clinical task such as clinical decision support.

XSL•FO
RenderX

Introduction

Background

Unstructured text from electronic health records (EHR) contains myriads of information that is not encoded in the structured part of EHRs, such as symptoms experienced by the patient. Structuring and managing symptom information is a major challenge for research owing to their complex and multidimensional nature. Extracting symptom information from clinical text is critical; for example, for phenotypic classification, clinical diagnosis, or clinical decision support [1-3]. More specifically, symptoms are crucial to the assessment and monitoring of the general state of the patient [1,4] and are critical indicators of quality of life for chronically ill patients [5,6]. Their evolution through time can be a string indicator of the patient's clinical status change. Finally, in the context of pandemic prevention, symptoms are used for syndromic surveillance [7,8] and patient characterization [9,10].

Using natural language processing (NLP) and machine learning to process and use clinical text for such applications has great potential [11-14]. Unfortunately, machine learning, and more specifically supervised machine learning, requires a great amount of labeled data to train a model, which is at the origin of the main bottleneck of model development [15]. Manually labeling data sets is extremely costly and time-consuming as multiple experts need to manually review and annotate several hundreds of clinical notes [13,16]. Moreover, the development of such a resource presents unique challenges as the text contains personal information, and access to such data is usually restricted.

Throughout the past years, shared resources such as Informatics for Integrating Biology and the Bedside (i2b2) have generated deidentified and annotated data sets for the development of NLP systems for specific tasks. Such resources remain limited, as most of the annotated data sets contain only hundreds to a few thousands of notes. Moreover, these data sets come from a limited number of institutions, making the development of an NLP system with such data unlikely to generalize to other institutions or other tasks.

To develop NLP systems and models that are transferable between multiple institutions and free of overfitting, a large amount of data needs to be available for training. To do so, alternatives to supervised machine learning have been explored, such as distant supervision, which seeks to include information from existing knowledge bases [17] or active learning, which involves human experts in the machine learning process [18-20]. One method in particular, weak supervision, is attracting increasing attention for the automatic generation of lower-quality labels for unlabeled data sets [21-25].

Objective

To address the lack of labeled data, we propose 2 alternatives to manual labeling for the generation of training labels for supervised machine learning with clinical text. The first approach takes advantage of the structured part of EHRs and uses diagnosis codes to derive training labels. The second approach uses weak supervision and data programming principles to derive training labels. We propose to apply the developed framework to the extraction of symptom information from outpatient visit progress notes of patients with cardiovascular diseases.

Extracting symptoms from clinical narratives is not a straightforward task as symptoms are often expressed in an abstract manner. A straightforward way of deriving labels from EHR would be to take advantage of their coded part and use the International Classification of Disease–10th revision–Clinical Modification (referred to as ICD-10, henceforth) codes. This approach has challenges, as demonstrated in multiple studies [2,9,26-30]. This is especially true if the target information is symptoms, as the corresponding ICD-10 chapter is typically used when a sign or symptom cannot be associated with a definitive diagnosis. Thus, their occurrence in EHR is very scarce and expected to be incomplete. Despite issues related to inaccuracy in ICD-10 coding, we propose to use such codes to label our training set, with the assumption that with sufficient training data, the poor quality of the labels will be balanced out. Although inaccurate and possibly biased, the use of ICD-10 data is considered standard in many classification studies involving clinical text [15,31-41]. Moreover, we propose to complement the use of ICD-10 codes with a weak supervision approach to derive labels. Weak supervision has gained a great amount of traction in the past years [21-25] as a response to the increased need for training data for machine learning. We used the Snorkel library [42] to combine a large number of clinical reports with noisy labeling functions and unsupervised generative modeling techniques to generate labels for our models. Finally, we test the models on external cohorts as a way to assess the bias and test the generalizability of the models.

We successfully demonstrate that by using a large number of notes for training, we can train a classification model able to recognize specific classes of symptoms using low-quality labels. The resulting model is independent of the prevalence of positive instances and is transferable to a different institution. We show that training our model on such pseudolabels results in a good predictive performance when tested on a data set containing gold labels.

Methods

Cohort Description

Our data set consisted of 20,009,822 notes from January 1, 2000, to December 31, 2016, for 134,000 patients with cardiovascular diseases from Stanford Health Care (SHC), collected retrospectively in accordance with the approved institutional review board protocol (IRB-50033) guidelines. Progress notes from outpatient office visits were selected. As the ICD-10 codes for symptoms were chosen for initial labels, encounters without R codes were discarded. Finally, short notes (ie, <350 characters) were also discarded. The final cohort contained 545,468 notes for 93,277 patients (Figure 1).

For prototyping purposes and to evaluate the effect of the training set size on the performance, subsets of the full cohort were created, leading to the following three data set sizes: I

(patients: 717/93,277, 0.77%), II (patients: 5611/93,277, 6.02%), and III (patients: 93,277/93,277, 100%). Patients were split into training, validation, and test sets using a 60:20:20 ratio. Table 1 provides a more detailed description of the data sets.

ICD-10 codes describing symptoms and signs involving the circulatory and respiratory systems were used to label the notes for the text classification task. The symptoms considered were only coded at the highest level of the ICD-10 hierarchy. The prevalence of the R codes was low, between 2% and 10% of positive instances (see Table S1 in Multimedia Appendix 1 for details).

Figure 1. CONSORT (Consolidated Standards of Reporting Trials) diagram for Stanford Health Care–electronic health record symptom extraction. Our full cohort consisted of 20 million notes and 134,000 patients. We selected progress notes from outpatient visits from encounters with International Classification of Disease–10th revision (ICD-10) codes from the chapter R. Notes <350 characters were discarded, yielding 545,468 notes for 93,277 patients.
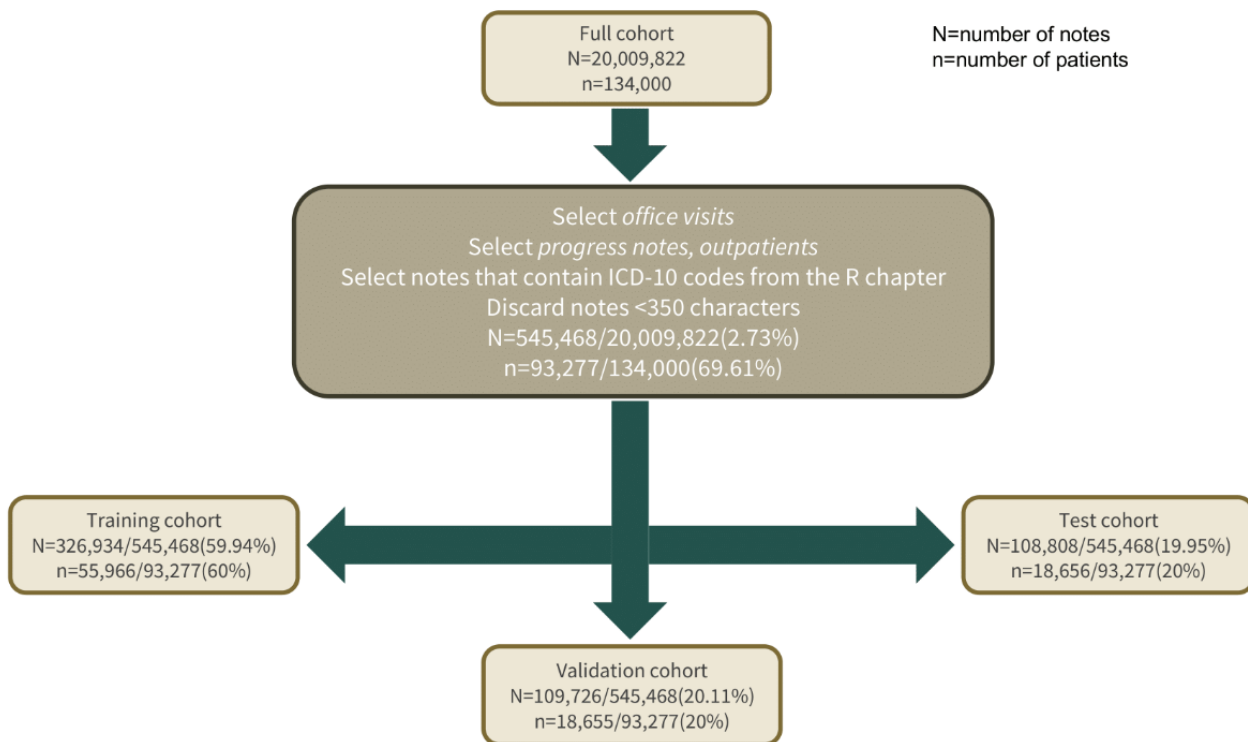
Table 1. Patient and note distribution for each data set considered in this study.

Data set	I[a] (N=717)	II[a] (N=5611)	III (N=93,277)	IV[b] (93,277)	V[c] (N=75.692)
Train set, n (%)	430 (59.9)	3360 (59.88)	55,966 (59.99)	55,966 (59.99)	38,381 (50.71)
Validation set, n (%)	143 (19.9)	1123 (20.01)	18,655 (19.99)	18,655 (19.99)	18,655 (24.65)
Test set, n (%)	144 (20.1)	1128 (20.10)	18,656 (20)	18,656 (20)	18,656 (24.65)
Age (years), mean (SD)	60 (23)	58 (23)	59 (23)	59 (23)	53 (23)
Gender, n (%)					
Men	306 (42.7)	2381 (42.43)	51,876 (55.61)	51,876 (55.61)	43,765 (57.82)
Women	410 (57.2)	3229 (57.55)	41,396 (44.38)	41,396 (44.38)	31,925 (42.18)
Unknown	1 (0.1)	1 (0.02)	5 (0.005)	5 (0.005)	2 (0.003)
Total notes, n (%)	4245 (100)	34,368 (100)	545,468 (100)	871,753 (100)	544,907 (100)
Train set	2480 (58.42)	20,500 (59.65)	326,934 (59.94)	653,219 (74.93)	326,373 (59.89)
Validation set	704 (16.58)	6698 (19.49)	109,726 (20.12)	109,726 (12.59)	109,726 (20.14)
Test set	794 (18.70)	6494 (18.89)	108,808 (19.95)	108,808 (12.48)	108,808 (19.97)

[a]Data sets I and II are subsets of data set III.

[b]Data set IV represents the hybrid data set of labeled and unlabeled notes considered for the weak supervision experiment.

[c]Data set V contains the set of unlabeled notes from IV.

Pipeline

We defined our task of extracting symptom information from clinical notes as a multi-class classification problem. Machine learning algorithms were trained to classify whether each input note contained a specific class of symptoms.

The proposed pipeline used a subset of the ICD-10 chapter containing symptoms, signs, and abnormal clinical and laboratory findings. The codes in this chapter are typically used when a sign or symptom cannot be associated with a definitive diagnosis. As their occurrence in EHR is expected to be incomplete, we assumed that the presence of a code is associated with the observation of the symptom, but the absence of a code cannot be associated with the absence of the symptom in question.

The full pipeline developed for this study is depicted in Figure 2. We obtained the raw clinical text and encounter data from the SHC database. The raw text was first preprocessed for standardization purposes. Then, the text was transformed into a numerical format (ie, featurization) so that it can be used as input features for our model training. Then, ICD-10 codes were extracted from the structured encounter data to use as labels. A multi-class classification model was then trained to predict the presence of symptoms in the text. Next, we propose a weak supervision labeling pipeline as an additional method for extracting labels for the downstream prediction task. For that additional part, notes that were initially discarded because of the lack of symptom codes in the encounter data were processed using an entity recognition model with the spaCy library [43] and labeled using a labeling model generated using the Snorkel package [42].

Figure 2. End-to-end pipeline developed for extracting pseudolabels out of an electronic health record (EHR) database and training a text classifier for recognition of presence or absence of symptoms. The approach leverages the structured part of EHR (International Classification of Disease–10th revision–Clinical Modification [ICD-10–CM] codes) and weak supervision to generate labeled training corpus. Three types of labels are used for the training: ICD-10–CM codes; noisy labels obtained by a weak supervision pipeline; and hybrid labels, containing both ICD-10–CM codes and noisy labels. Two machine learning algorithms are considered: random forest and logistic regression. Four featurization methods are considered: bag-of-words (BOW), term frequency–inverse document frequency (TF-IDF), continuous BOW (CBOW), and paragraph vector–distributed BOW (PV-DBOW). LF: labeling function.
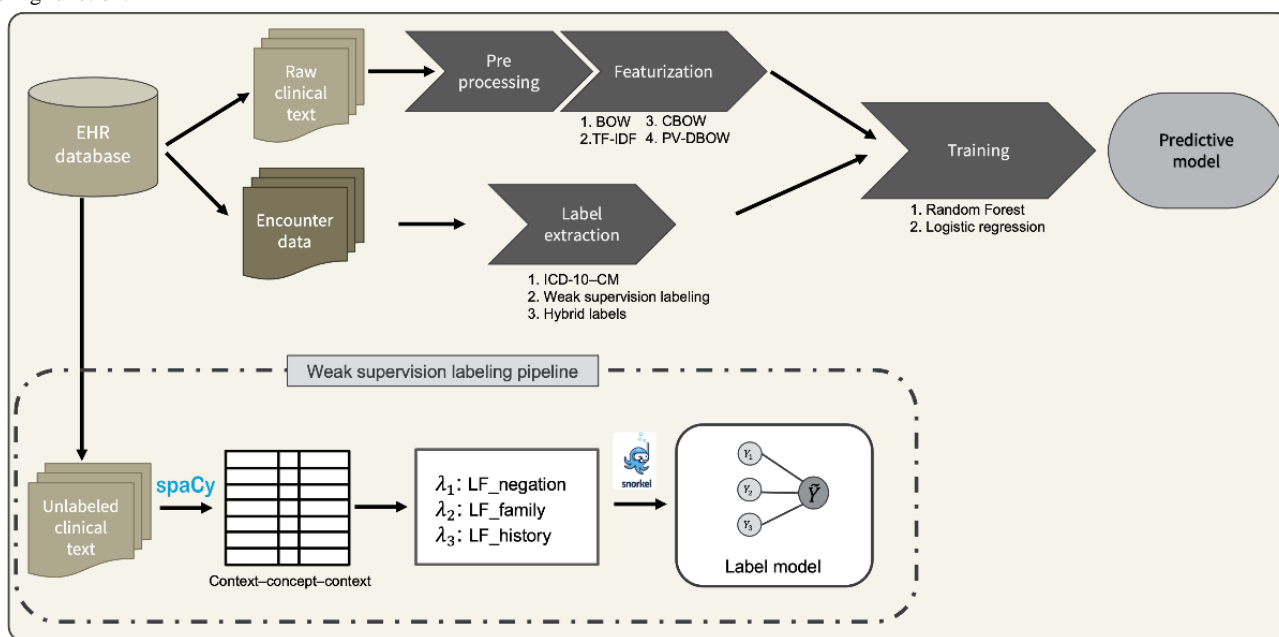
Preprocessing

To facilitate machine learning techniques, the clinical notes were standardized in the following manner: special characters and numbers were removed; the text was transformed into lower case only; frequent words (eg, the, as, and thus) often denoted as stop words were removed, except negative attributes such as no or not; next, each note was standardized using the Porter stemming algorithm; and finally, the text was tokenized into individual words. Sectioning of the notes was not performed; thus, the entire note was included in the featurization step.

Featurization

In this report, we evaluated the following approaches for featurization of the clinical notes. The first method, bag-of-words (BOW), is a simple yet effective method to represent text data for machine learning and acts as a baseline. In this method, the frequency of each word is counted, yielding a vector representing the document. As each word represents a dimension of the document vector, the size of the latter is proportional to the size of the vocabulary used. As words are represented by their document frequency, the resulting document vector does not contain any syntactic or contextual information.

Next, we used term frequency–inverse document frequency (TF-IDF), a weighting scheme, in addition to BOW whereby word frequencies from BOW are weighted according to their IDF. This reweighting of the frequencies dampens the effect of extremely frequent or rare words.

Next, we used the continuous BOW (CBOW; also referred to as *word2vec*) algorithm [44]. CBOW is an algorithm that generates word vectors based on a prediction task via a neural

network. The output of such a network is an embedding matrix that is used to encode each word into a specific vector. The embedding matrix used in this project was trained on biomedical text (PubMed and Medical Information Mart for Intensive Care–III [MIMIC-III]) by Zhang et al [45]. Word vectors were generated using these pretrained embeddings and then averaged to yield a single document vector representing the entire note. As a result, the document embedding vector was of dimension 200.

Finally, the paragraph vector–distributed BOW (PV-DBOW; also referred to as *doc2vec*) [46], an extension of CBOW to paragraphs, was used to add some syntactic knowledge in the encoding of each document. The vector size for the document was 300 and was independent of the corpus size.

Weak Labeling

To address the problem of a lack of labels for EHR-based supervised learning, a weak supervision pipeline using the Snorkel package [42] was implemented. Weak supervision allows us to create a set of noisy labels for an unlabeled data set. The noisy labels are generated using a set of *labeling functions*, namely, a set of heuristic rules.

For this project, we implemented labeling functions based on pattern recognition applied to a 20 token–context window (10 tokens before and 10 tokens after the target term) to determine the negation, temporality, and experiencer of the target symptom. We used the publicly available *clinical event recognizer* base terminology [47] to match our context window with negative expressions, historical expressions, and family mentions. If a mention is matched within the context window of a given term, it is labeled accordingly: *absent* if negative expression is matched, *history* if historical expression is matched, and *family* if family mention is matched. Target symptoms that were positive, experienced by the patient, and not part of the past medical history were labeled positive. Occurrences deviating from this pattern were labeled negative.

Symptom recognition was performed using a ScispaCy [48] pipeline trained to recognize biomedical entities. The process of extracting the presence or absence of symptoms belonging to the R00-R09 categories was implemented as follows: the full clinical note is processed with spaCy [43] using the entity recognition model from the ScispaCy library, trained on BioCreative V Chemical Disease Relation corpus, a corpus of 1500 PubMed articles annotated for chemicals, disease, and chemical–disease interactions [49] (en_ner_bc5cdr_md [48]). As we were classifying the notes using only the 3 characters categories of the ICD-10 codes, each entity that was tagged needed to be associated to its corresponding category. For that purpose, we normalized them to the concept unique identifiers from the unified medical language system with the highest similarity score. This allowed us to group each entity to their corresponding ICD-10 category (see Table S2 in Multimedia Appendix 1 for a list of concept unique identifiers). Then, the labeling functions defined earlier were used to generate noisy labels, which can finally be used to train a machine learning model.

Modeling

The input features were used to predict a set of symptoms related to abnormalities in the circulatory and respiratory systems (ICD-10 codes R00-R09). The problem was approached as a text classification task using a subset of the ICD-10-R codes for the class labels. The classes are not mutually exclusive; therefore, a *one-versus-all* classification was chosen. We compared two classification algorithms for this task, namely random forests [50] and logistic regression [51]. We only report the results obtained with 100 estimators for the random forest and the limited-memory Broyden–Fletcher–GoldfarbShanno solver. The detailed parameters used for each model are provided in the Multimedia Appendix 1.

Performance Evaluation

We used the following classification metrics to evaluate each model: recall, F1 score, and average precision score. We also computed the receiver operating characteristic (ROC) curves and precision-recall curves. Owing to the class imbalance, we gave more importance to the precision-recall curve. For example, in the case of *hemorrhage from respiratory passages* class of symptoms (R04), the positive instances represent only approximately 1% of the data points. We also considered computation time and memory requirements as important metrics to determine the best classification model. Given the size of our data set, an efficient implementation was of paramount importance for the success of our predictive model.

External Validation

To assess the impact of training the model on low-quality labels, the models were tested on an external data set developed for symptom extraction by Steinkamp et al [52]. Their work provides an open-source annotated data set for symptom extraction. The notes were 1008 deidentified discharge summaries from the i2b2 2009 Medication Challenge [53]. The set of notes was annotated by 4 independent annotators for all symptom mentions, whether *positive*, *negative,* or *uncertain*. To benchmark our study, we chose three classes of symptoms that were both present in our study and in the annotated data set of Steinkamp et al [52], namely cough (R05), abnormalities of breathing (R06), and pain in throat and chest (R07). As the annotations were performed at the *mention* level but our study was performed at the *note* level, a majority voting algorithm was chosen to assess the note-level polarity of the symptom mention to generate note-level labels. On the basis of the SHC experiments, only models showing the best promise in terms of predictive performance were chosen for this step. More specifically, models trained with the logistic regression algorithm using TF-IDF and PV-DBOW features were chosen for the external validation.

Results

Logistic Regression Performs Better Than Random Forest for Predicting the Presence of Symptoms in Outpatient Progress Notes

Outpatient progress notes collected from January 1, 2000, to December 31, 2016, from the SHC EHR database were used to train a text classifier to extract symptoms related to

abnormalities in the circulatory and respiratory systems (Figure 1). Two machine learning algorithms were considered, namely random forest and logistic regression. The models were first built on a subset of the cohort for prototyping purposes (Table 1: data set I). Random forest showed poor predictive performance, with no or few positive instances predicted (Figure 3). Without exception, logistic regression outperformed random forest for all the considered data set sizes (Figure 4). Use of TF-IDF features to predict the presence of symptoms in the notes led to the best overall performance (Figure 5).

Figure 3. Histogram of predicted probabilities for the presence of the cough symptom (R05) in the outpatient progress note for data set I, with a comparison between probabilities predicted by logistic regression (LR) and random forest (RF) for term frequency–inverse document frequency (TF-IDF) and paragraph vector–distributed bag-of-words (PV-DBOW) feature extraction methods.
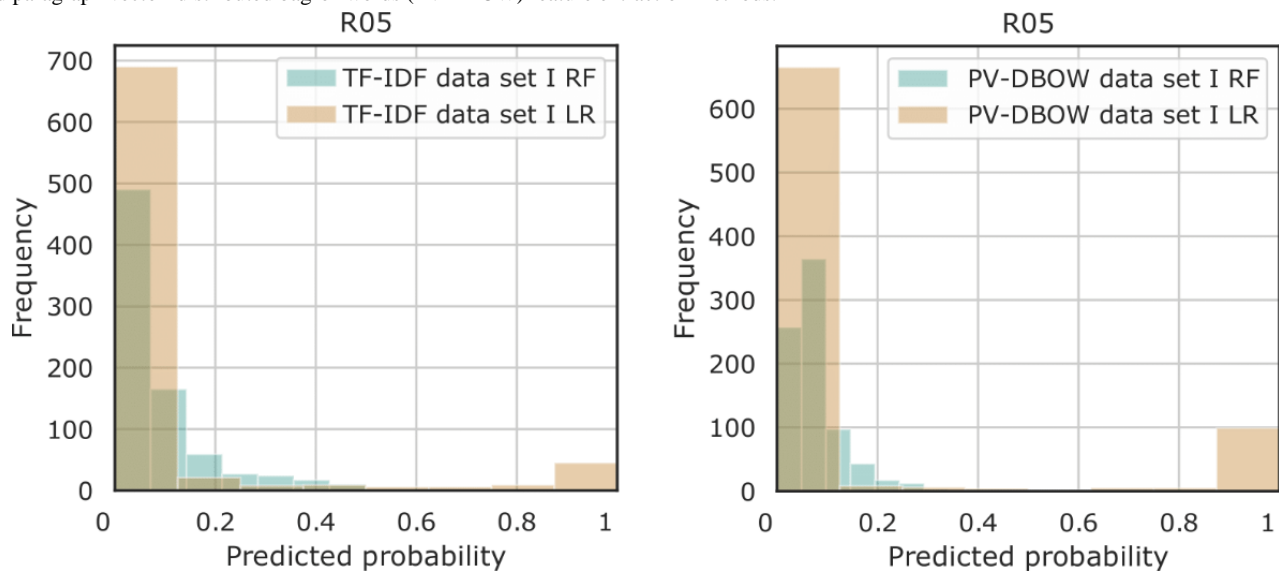
Figure 4. Summary of performance metrics averaged over all codes for all four considered feature extraction methods (bag-of-words [BOW], term frequency–inverse document frequency [TF-IDF], continuous BOW [CBOW], and paragraph vector–distributed BOW [PV-DBOW]). AUROC: area under the receiver operating characteristic curve; LR: logistic regression model; RF: random forest model.
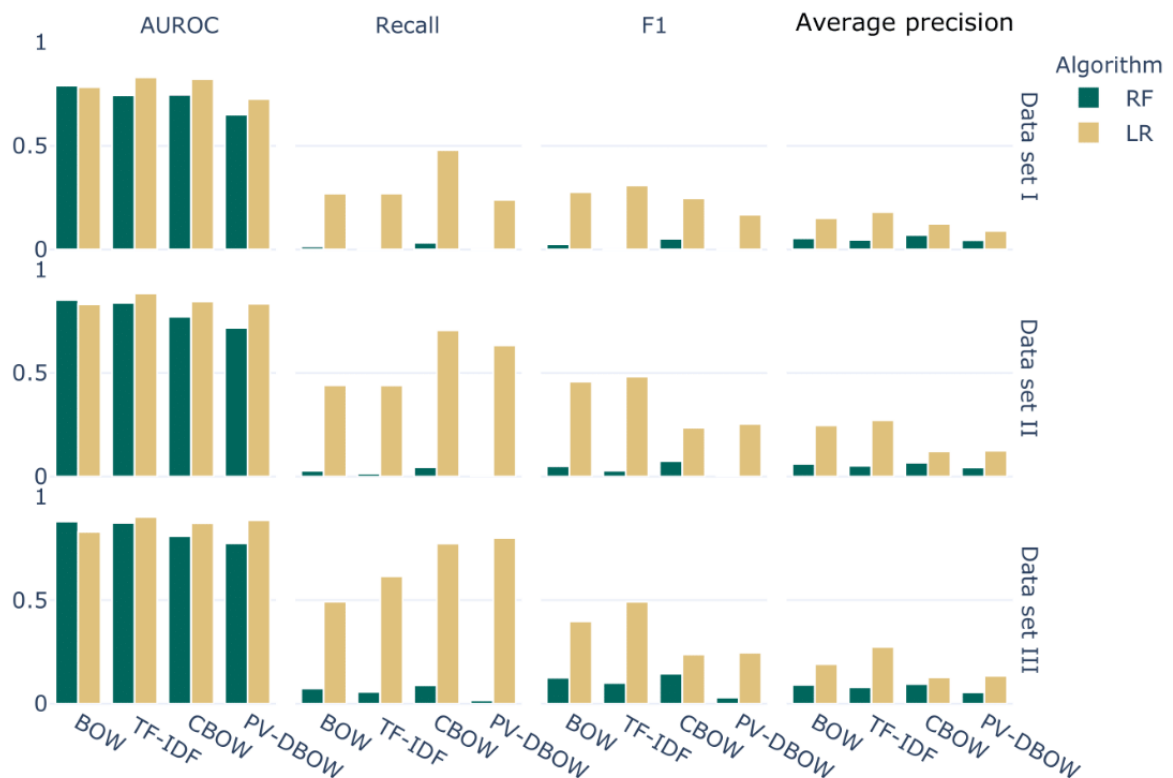
Figure 5. Receiver operating characteristic and precision-recall curves for the prediction on the test set (data set I described in Table 1) of presence of cough (R05) symptoms from outpatient progress notes using logistic regression (LR) with 4 feature extraction methods. BOW: bag-of-words; CBOW: continuous BOW; lbfgs: limited-memory Broyden–Fletcher–GoldfarbShanno solver; PV-BOW: paragraph vector–distributed BOW; TF-IDF: term frequency–inverse document frequency.
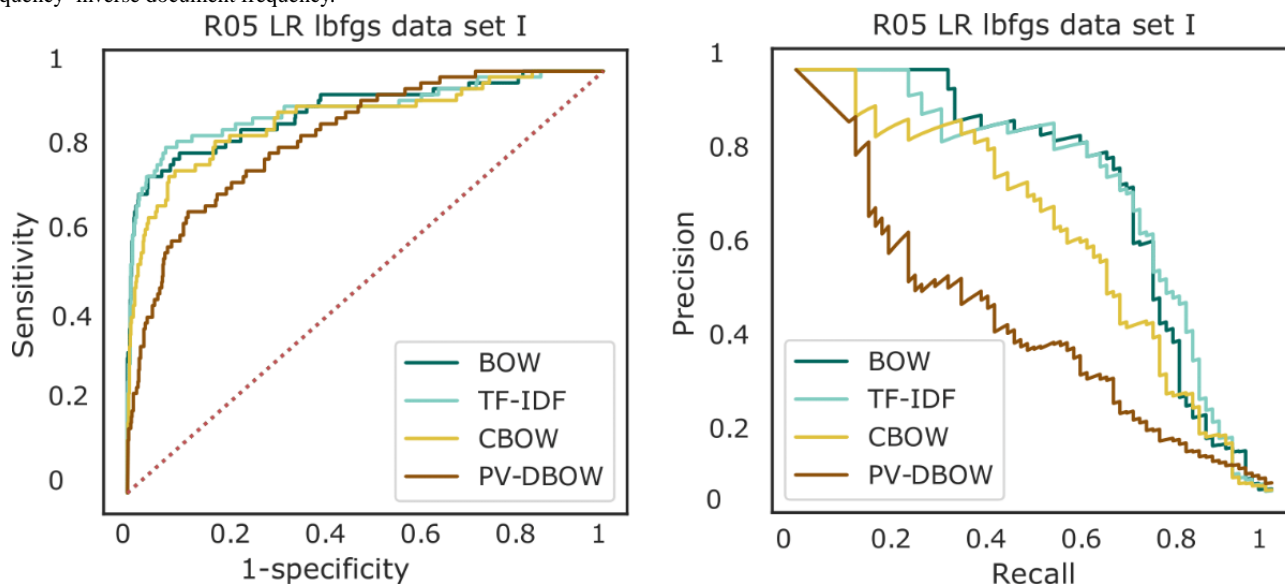
Embedding-Based Methods Perform Better With Increasing Data Set Size

To demonstrate that increasing the size of the training set significantly improves the performance of deep learning–based embedding methods, the classification task was performed on 3 different data set sizes, ranging from 0.75% (700/93,277) of patients to 100% (93,277/93,277) of patients (Table 1).

For all codes, the performance (area under the ROC [AUROC] curves and area under the precision-recall curves) of PV-DBOW features with logistic regression drastically improved with the size of the training set. For TF-IDF features also, there was a slight improvement, but it was less pronounced (Figure 6). More importantly, we observed that when increasing the size of the training set, the low prevalence of the symptoms does not affect the performance of embedding-based features (CBOW and PV-DBOW; Figure 7). Next, although the performance obtained with TF-IDF features was high, the computational performance was drastically affected by the increasing size of the training set. It takes 2 minutes and 1.6 GB of memory to train the model with PV-DBOW features, whereas the model with TF-IDF features requires 2.3 GB of memory and takes almost 3 hours (Table 2).

Figure 6. Comparison of receiver operating characteristic (left column) and precision-recall (right column) curves for the prediction of presence of cough (R05), abnormality of breathing (R06), and pain in throat and chest (R07) classes of symptoms from outpatient progress notes using logistic regression (LR) with the limited-memory Broyden–Fletcher–GoldfarbShanno (lbfgs) solver on data set I, data set II and data set III with term frequency–inverse document frequency (TF-IDF) and paragraph vector–distributed bag-of-words (PV-DBOW) features.
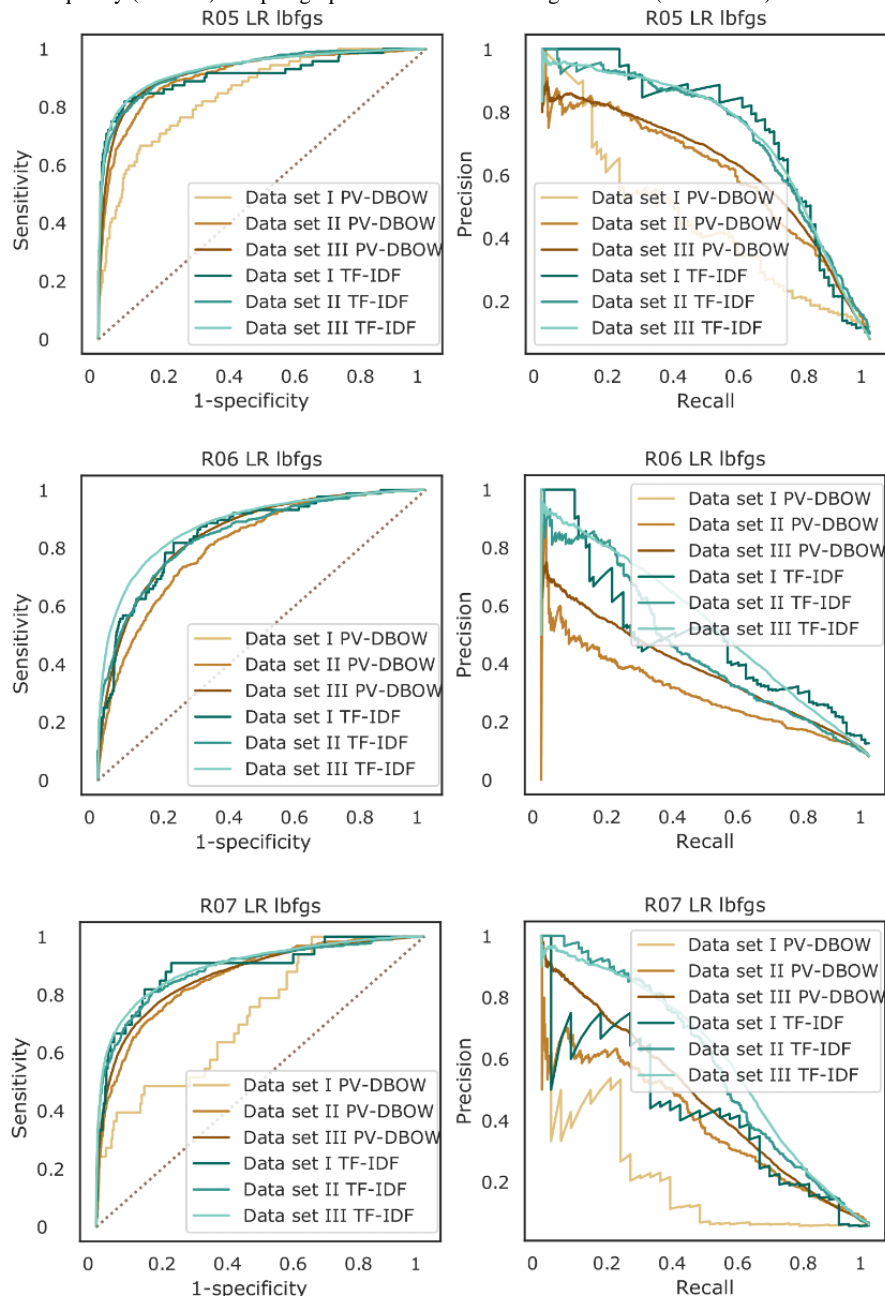
Figure 7. Recall scores as a function of the symptom prevalence in 3 considered data sets for all the features. BOW: bag-of-words; CBOW: continuous BOW; PV-BOW: paragraph vector–distributed BOW; TF-IDF: term frequency–inverse document frequency.
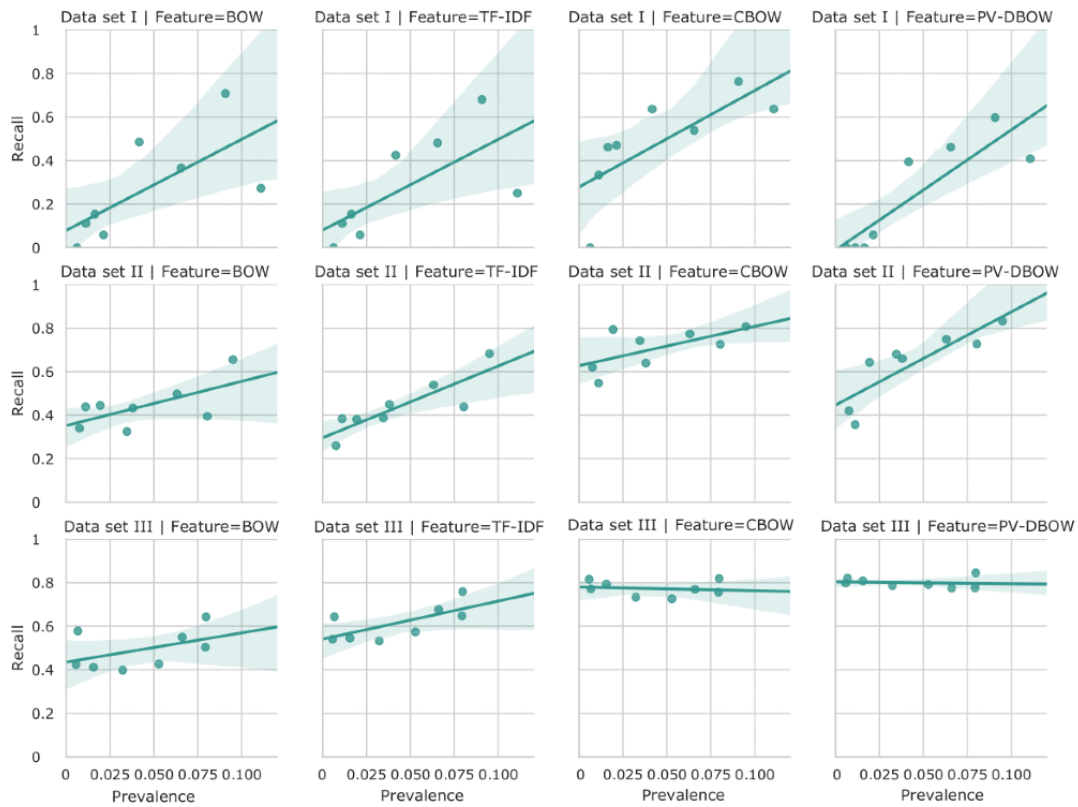
Table 2. Computational resources used for each classifier by feature type for data sets II and III.

Feature type and data set	Random forest		Logistic regression	
	Memory, MB	Run time, hours:minutes:seconds	Memory, MB	Run time, hours:minutes:seconds
BOW[a]				
II	310	00:04:10	340	00:21:35
III	3500	07:22:02	3400	23:17:20[b]
TF-IDF[c]				
II	310	00:04:15	270	00:03:04
III	3400	06:37:04	2300	02:47:30
CBOW[d]				
II	193	00:03:02	180	00:01:17
III	1700	01:21:11	1700	00:16:36
PV-DBOW[e]				
II	170	00:03:35	89	00:00:34
III	1100	01:41:18	1600	00:02:13

[a]BOW: bag-of-words.

[b]No convergence after 100,000 iterations.

[c]TF-IDF: term frequency–inverse document frequency.

[d]CBOW: continuous BOW.

[e]PV-DBOW: paragraph vector–distributed BOW.

Enriching the Training Set With Weak Labels Enhances the Performance Further

The original cohort contained many notes that do not contain ICD-10 codes from the R chapter, leading to a substantial reduction in the number of notes available to train our model. Indeed, an additional 1,290,170 notes from the patients included in our cohort did not contain any ICD-10 code for symptoms.

To use these notes, they were processed using a weak supervision approach to determine the presence or absence of symptoms belonging to the R00-R09 categories. Then, the weakly labeled notes were added to data set D for training the classifier (ie, data set IV). For comparison, we also trained a model using only the weakly labeled notes (ie, data set V). Then, the 2 models were tested on test set III with ICD-10 codes for

labels. The weak labeling model was also applied to the test set to extract weak labels for testing. Given the poor scaling performance of TF-IDF features compared with that of PV-DBOW, this experiment was performed solely with the PV-DBOW features.

Figure 8 shows the difference in performance between the enriched data set (IV) and the baseline data set (III). Overall, the recall score improved by 3.8%. However, the AUROC score was reduced by 2.1%. This decrease in the AUROC score can be attributed to the number of false-positive predictions. As the model was trained on mixed labels (ICD-10 and weak labels) but tested on ICD-10 codes, such increase in predictions flagged as false positives was expected. However, treating the weak labels as *true* labels for the test set led to an increase in recall score by 17.7% and an increase in AUROC score by 3.7%.

Figure 8. Performance metrics differential for the weak labeling experiment. Delta H represents the score difference between the hybrid data set IV and the baseline data set III (score [IV]–score [III]). Delta W represents the score difference between the weakly labeled data set V and the baseline data set III (score[V]–score [III]) The left panel shows the score calculated using International Classification of Disease–10th revision–R (ICD-10–R) codes for labels and the right panel shows the score calculated treating the weak labels (WL) as true labels in the test set. AUROC: area under the receiver operating characteristic curve.
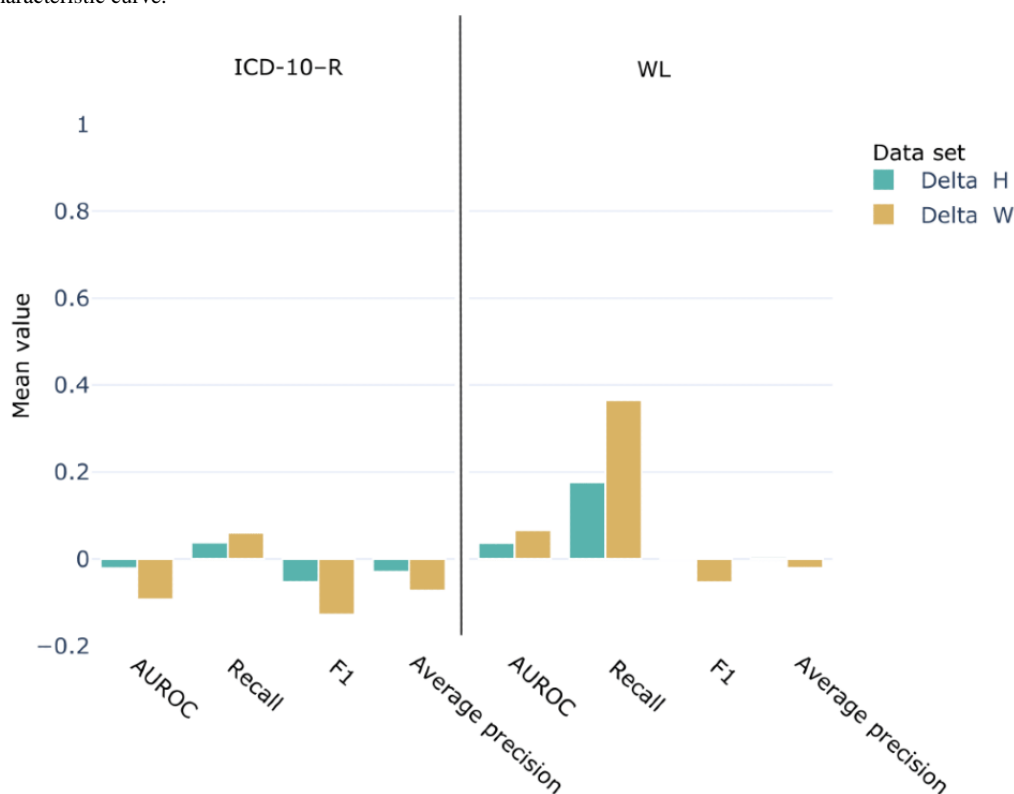
Use of only weakly labeled notes for training (data set V) and testing on ICD-10 labels led to a 6% increase in recall score and a 9.3% decrease in the AUROC score. Finally, using the weak labels as *true* labels for the test set, the weakly labeled notes performed 36.6% (recall) and 6.6% (AUROC) better than the baseline data set.

Embedding-Based Features Perform Better Than TF-IDF Features on an External Validation Set

We selected a set of 56.65% (571/1008) notes from the i2b2 2009 challenge annotated for symptom extraction [52] containing mentions of symptoms of cough (R05), abnormalities

of breathing (R06), and pain in throat and chest (R07). The logistic regression models trained on data set III using TF-IDF and PV-DBOW features were used to predict the presence of the 3 classes of symptoms.

Overall, the model trained with PV-DBOW features performed well when used to predict symptoms from the i2b2 notes. Figure 9 shows the difference in scores between the i2b2 data set and the baseline data set III trained using TF-IDF and PV-DBOW features for the set of 3 selected classes of symptoms. For R06 and R07, PV-DBOW, recall, and AUROC scores were within the range of the scores obtained when tested on the SHC notes.

However, the F1 and average precision scores were >40 points better on the i2b2 notes. On the other hand, the model trained with TF-IDF features performed poorly. The recall and AUROC scores were 20 to 30 points lower than when tested on the SHC notes. The F1 score was similar to that obtained with the SHC notes. However, the average precision was almost 30 points higher than that of the SHC notes (Figure 9). For both PV-DBOW and TF-IDF features, the performance of the symptom *cough* decreased when tested on the i2b2 set compared with the SHC notes.

Finally, the models trained with the hybrid labels and weak labels using the PV-DBOW features were also tested on the i2b2 notes. For both models, the recall and AUROC scores were within the range of those obtained with the SHC notes. However, the F1 and average precision scores were approximately 50 points higher than when tested with the SHC notes, reinforcing the conclusion that even though the models were trained on pseudolabels, they still perform well when tested on gold labels (Figure 10). Typically, recall performed better when hybrid or weak labels were used for training than when ICD-10 codes were used. Similar to the use of ICD-10 codes as labels, the performance for R05 decreased for the i2b2 notes.

Figure 9. Performance metrics differential for the external validation set. The score has been calculated as the difference between the score obtained on the external validation set and the baseline data set III (score [Informatics for Integrating Biology and the Bedside]–score [Stanford Health Care]). Term frequency–inverse document frequency (TF-IDF) represents the logistic regression model trained with TF-IDF features. Paragraph vector–distributed bag-of-words (PV-DBOW) represents the logistic regression model trained with PV-DBOW features. International Classification of Disease–10th revision–R codes have been used as reference labels to compute the metrics. AUROC: area under the receiver operating characteristic curve.
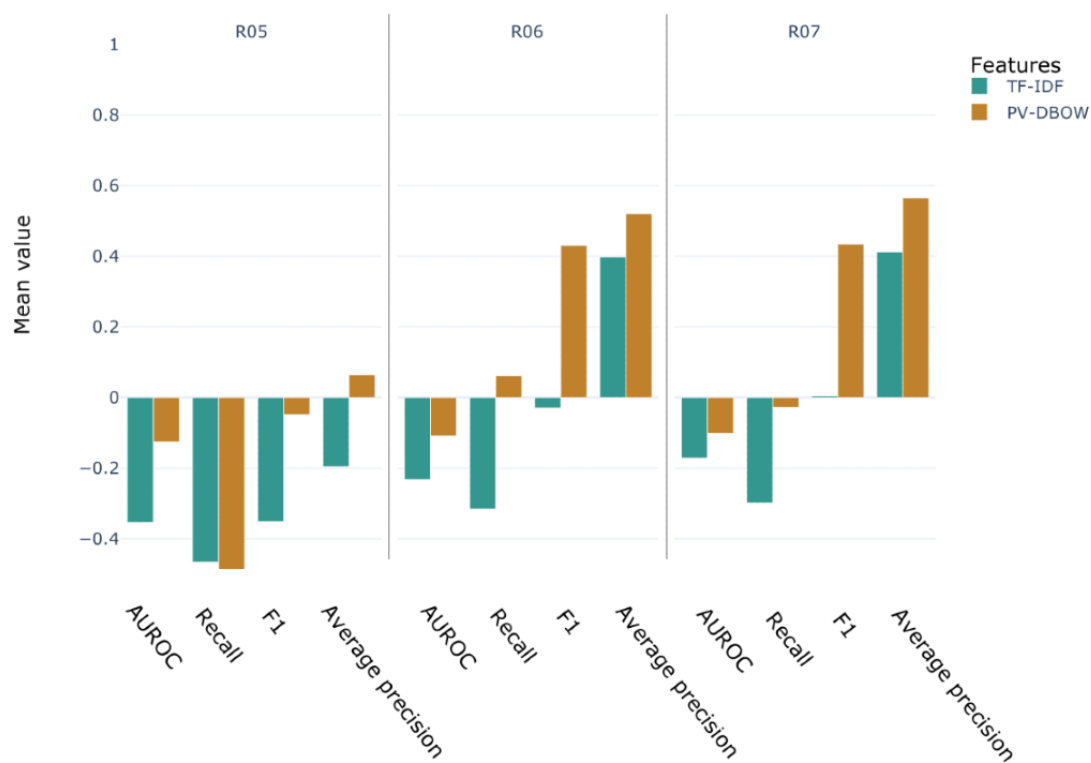
Figure 10. Performance metrics differential for the external validation set. The validation was performed for three models using paragraph vector–distributed bag-of-words features only, trained using different labels: International Classification of Disease–10th revision–R, the weak labels, and the hybrid labels. The score differences are computed relative to the baseline data set III (score [Informatics for Integrating Biology and the Bedside]–score [Stanford Health Care]). AUROC: area under the receiver operating characteristic curve.
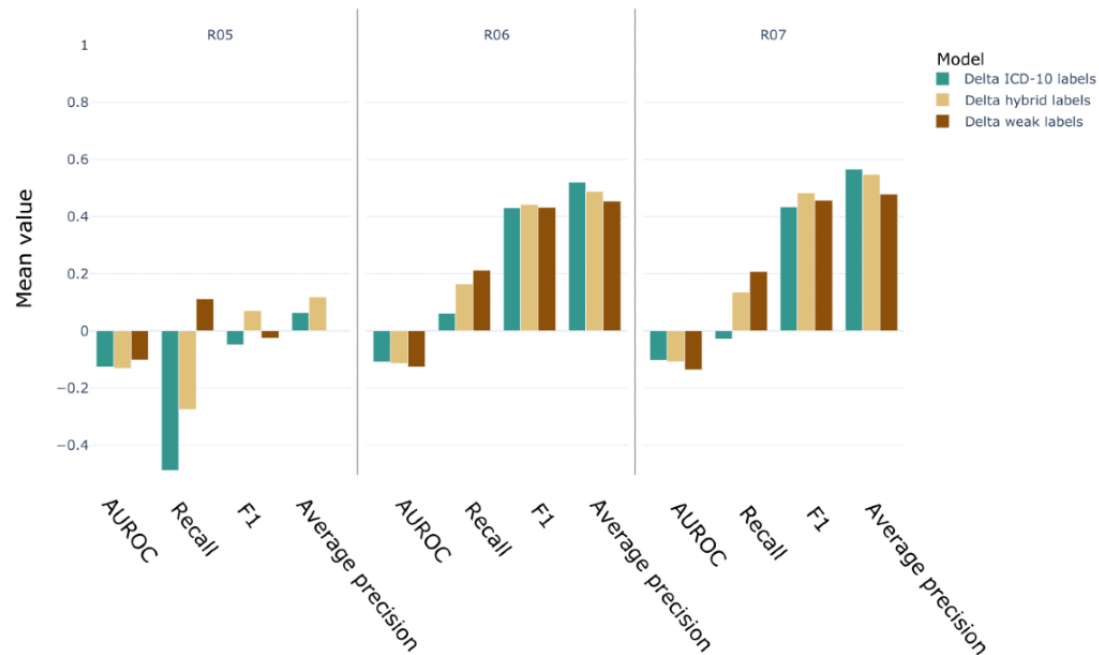
Analysis of Misclassified Cases

To illustrate that despite the low quality of training labels used, the classification models were able to correctly classify notes, we show a few examples of the presence of abnormality of breathing symptoms in Figure 11. Snippets (A) and (E) show examples where the predictions were flagged as false positive but turned out to be true-positive cases. Snippets (B) and (C) show 2 examples that were flagged as false negative; however, when reading the note, the symptom was clearly absent (historical for (B) and negated for (C)). Finally, snippet (D) shows an example that was correctly predicted only when embedding features were used.

Figure 11. Snippet examples of mislabeled notes for R06 class of symptoms. ICD-10: International Classification of Disease–10th revision; NEG: negative; POS: positive; WL: weak labels.


```
A. Labeled NEG (ICD-10)-predicted POS

   'Nursing Note'''Health Maintenance Due Topic Date Due PNEUMOCOCCAL
   VACCINE 65 AND OVER [DATE] Chief Complaint Patient presents with Rib
   Pain Shortness Of Breath Foot […]

B. Labeled POS (ICD-10/WL)-predicted NEG
   'Clinic Visit[…] chief complaint of palpitations, as well as aortic
   and mitral valve disease. Since our last visit, I had recommended
   doing a stress test as she was getting some dyspnea on exertion when
   walking up hills.[…]. She has been noting palpitations happening at
   least once or twice a day, lasting for several seconds. Otherwise,
   no other complaints on a 14-point review. […]

C. Labeled POS (ICD-10)-predicted NEG
   'Progress Notes'''Samaritan Internal […] ROS: […] The patient denies
   hemoptysis, dyspnea or wheezing. No edema, palpitations, chest pain
   or SOB. The patient denies abdominal or flank pain, […]

D. Labeled POS, predicted POS only with embedding methods

   'Progress Notes'''CARDIOLOGY PROGRESS NOTE […] 3 or 4 days ago she
   started feeling SOB. To her, it feels like an asthma exacerbation.
   Her SOB is exertional. She has also been propping herself up to
   sleep at night. She feels like she can\'t breathe if she lies
   flat....[…]

E. Labeled NEG (ICD-10)-predicted POS

   'Progress Notes'''[…] who presents for evaluation of cough and
   congestion x 5-6 days. At onset tightness in chest and mild, dry
   cough. […] Last night was coughing more. Not as deep but cannot
   catch breath when coughing. Today when doing her daily back
   exercises she felt out of breath.
```

Discussion

Principal Findings

We trained *one-versus-all* multi-label classification models using four featurization methods, namely BOW, TF-IDF, CBOW, and PV-DBOW, to predict the presence of signs and symptoms related to abnormalities in the circulatory and respiratory systems. The challenging lack of labels for training such models was addressed using 2 label extraction strategies. First, we extracted labels based on a subset of ICD-10 codes from EHR encounter data. This approach yielded good predictive performance, as evidenced by external validation. Relying on the coded part of EHR to extract training labels leaves a large part of progress notes untouched, as ICD-10 codes for symptoms are rarely used. The second approach we used was a method to extract training labels by leveraging clinical named entity recognition and a weak supervision pipeline. This approach not only allowed us to make use of a much larger set of notes for training but also significantly improved the predictive performance, both on an SHC test set and an external validation set.

Although TF-IDF features yielded the best performance overall (Figure 4), the size of the feature vector is the size of the corpus,

leading rapidly to intractable size and computational inefficiency when the corpus size increased (Table 2), whereas embedding methods such as CBOW and PV-DBOW led to a fixed feature vector length, independent of the training corpus size. The main computing cost in such an approach lies in the pretraining of the embedding vectors, which must be performed only once. Training a classifier on any data set size led only to a minor increase in computational cost, making this approach more desirable.

Unfortunately, the results on a small training set were not satisfactory as these types of models are known to be extremely data hungry. The performance is expected to be more reasonable with larger data set sizes. We observed this in our experiments; when the training set size was increased, the performance also increased significantly. For example, the most notable performance improvement was observed for the recall, which increased from 0.25 to 0.8 for PV-DBOW features (Figure 4). This is important because when predicting the presence or absence of symptoms, minimizing the false-negative rate is desirable. Moreover, owing to the nature of our training labels, the absence of an ICD-10 code does not mean the absence of the symptom, whereas the presence of the code more likely signifies the presence of the symptom. Moreover, the effect of the low prevalence of some codes on the performance became

negligible with increasing data set size and the use of PV-DBOW features, suggesting that the use of a resampling method is not necessary if training on larger data sets (Figure 6).

Next, enriching the largest data set with unlabeled notes using a weak supervision approach for labeling yielded an overall gain in performance. This result not only suggests that more is better but also points to the conclusion that the use of ICD-10 codes as labels to extract the presence of symptoms from clinical notes can be improved by using weak labeling pipelines to label previously unlabeled notes. Indeed, external validation of our models showed a large increase in performance of the PV-DBOW features. We attribute this gain to the quality of labels in the external validation data set, resulting in a drop in false-positive predictions. This experiment also suggests that although the quality of the labels used to train the models was not optimal, the model was still able to learn enough to reliably predict the presence of symptoms. On the other hand, the poor performance of the TF-IDF features suggests that the high performance observed on the SHC notes might be owing to overfitting of the features rather than a good predictive power. However, the increase in average precision suggests that the false-positive rate is reduced owing to the higher quality of the labels. Although TF-IDF seems to work well within one context, it is likely to fail when testing at other sites.

It is worth noting that the performance for cough symptoms (R05) decreased significantly when tested on our external validation data set. The causes for such a drop have not been investigated, but Figure 10 offers some hints about a labeling issue. Indeed, the recall score performed poorly when using the model trained with ICD-10 codes as labels but increased when using the weak labels as ground truth for training.

The automatic classification of clinical text into specific ICD codes is a common task, and various state-of-the-art models have been developed over the years. Although our objective is different, it is worth comparing our classification results with some of the available work. Moons et al [54] recently compared multiple state-of-the-art models for ICD coding of clinical records, using public data sets encoded with both ICD-9 (MIMIC-III [55]) and ICD-10 (CodiEsp [56]). They reported micro- and macro-F1, micro-AUROC, and Precision@5 for multiple subsets of MIMIC-III and CodiEsp using multiple deep learning architectures. As they did not report recall or the prevalence of each class, a direct comparison with our work is difficult. However, it is worth noting that the best-performing model on the MIMIC-III data set (using ICD-9 codes) yields a macro-F1 of 64.85. Their best-performing model on CodiEsp (using ICD-10 codes) yields a macro-F1 of 11.03. Our macro-F1 of 24.66 falls in between these values, suggesting that our performance lies within the range of some of the best-performing deep learning models available.

We note that although we are using a data set containing gold standard annotations, a direct comparison with previous results from Steinkamp et al [52] is not possible. Both experiments are fundamentally different. Our objective was to lay out strategies to generate training labels for a symptom classification task and demonstrate that if sufficient training data are provided, such strategies will yield good predictive performance. We did not aim to extract all symptoms from the notes or create new named entity recognition models. The use of the external data set, labeled by Steinkamp et al [52], was meant to show that (1) our models, although trained on SHC data, perform well on another institution's data and, (2) considering that our models were trained on pseudolabels, they performed well on a test set containing gold labels.

Recent work has also seen the rise in transformers for NLP tasks. Although these methods are gaining popularity, the adaptation of such language model to the clinical use case is not straightforward. First, transformer models usually have a relatively short fixed maximum input length (eg, 412 tokens for bidirectional encoder representations from transformers [BERT]–based models). Clinical notes in general, and progress notes in particular, tend to be much longer than that (eg, in our case, the note length is closer to a couple of thousands of tokens). Moreover, transformer-based models trained on open domain text are not suitable for clinical text and must be fine-tuned to maximize performance. Although some BERT adaptations for the clinical domain have been released recently (eg, ClinicalBERT [57], BioBERT [58], or BlueBERT [59]), these publicly available models might not be suitable for the task at hand. Reasons why BERT-based models might not be suitable include attention dilution and the use of subword tokenization rather than word-level tokenization [60]. Finally, finding the best embedding method for note classification was outside the scope of our study. For these reasons, we did not include transformers in our comparison.

Conclusions

In this study, we introduced 2 methods to extract labels from EHR data sets for the training of a classifier for clinical notes. Multiple featurization methods were investigated, showing that PV-DBOW is clearly superior in terms of transferability and scaling. Although the use of ICD-10 codes present in the encounter data is a simple way of extracting training labels, the poor accuracy of the coding leads to less accurate models. Using a weak labeling pipeline to extract such labels yields improved performance and allows for the use of more notes as we are not relying on the presence of codes. Both approaches have been validated with an external set of notes containing gold labels, which showed the superiority of the weak labeling approach. Using ICD-10 codes for initial labels, we grouped a wide variety of signs and symptoms under the same label, learning classes of symptoms rather than specific symptoms. For example, R06 (abnormalities of breathing) covers a variety of breathing abnormalities; for example, dyspnea, wheezing, or hyperventilation. Such granularity in the symptoms is beyond the scope of this study and thus has not been investigated. However, the good performance of the weak labeling pipeline suggests that such an approach to generate more granular labels (eg, to distinguish between wheezing and shortness of breath in the R06 category) could be used. Moreover, the nature of the *one-versus-all* approach allows us to add a new category without having to retrain our model on all labels. Finally, the good performance and computational efficiency of the PV-DBOW features with logistic regression model would make such an expansion of the model computationally cheap.

Data and Code Availability

Protected Health Information restrictions apply to the availability of the Stanford Health Care clinical data set presented here, which were used under institutional review board approval for use only in this study and thus are not publicly available. The code can be made available upon request.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Supplementary tables.
[DOCX File , 43 KB - medinform_v10i3e32903_app1.docx]

References

1. Koleck TA, Dreisbach C, Bourne PE, Bakken S. Natural language processing of symptoms documented in free-text narratives of electronic health records: a systematic review. J Am Med Inform Assoc 2019 Apr 01;26(4):364-379 [FREE Full text] [doi: 10.1093/jamia/ocy173] [Medline: 30726935]
2. Forbush TB, Gundlapalli AV, Palmer MN, Shen S, South BR, Divita G, et al. "Sitting on pins and needles": characterization of symptom descriptions in clinical notes". AMIA Jt Summits Transl Sci Proc 2013;2013:67-71 [FREE Full text] [Medline: 24303238]
3. Adnan K, Akbar R, Khor S, Ali A. Role and challenges of unstructured big data in healthcare. In: Data Management, Analytics and Innovation. Singapore: Springer; 2020:301-323.
4. Koleck TA, Tatonetti NP, Bakken S, Mitha S, Henderson MM, George M, et al. Identifying symptom information in clinical notes using natural language processing. Nurs Res 2021;70(3):173-183. [doi: 10.1097/NNR.0000000000000488] [Medline: 33196504]
5. Luo X, Gandhi P, Storey S, Zhang Z, Han Z, Huang K. A computational framework to analyze the associations between symptoms and cancer patient attributes post chemotherapy using EHR data. IEEE J Biomed Health Inform 2021 Nov;25(11):4098-4109. [doi: 10.1109/jbhi.2021.3117238]
6. Forsyth AW, Barzilay R, Hughes KS, Lui D, Lorenz KA, Enzinger A, et al. Machine learning methods to extract documentation of breast cancer symptoms from electronic health records. J Pain Symptom Manag 2018 Jun;55(6):1492-1499 [FREE Full text] [doi: 10.1016/j.jpainsymman.2018.02.016] [Medline: 29496537]
7. Mandl KD, Overhage JM, Wagner MM, Lober WB, Sebastiani P, Mostashari F, et al. Implementing syndromic surveillance: a practical guide informed by the early experience. J Am Med Inform Assoc 2004;11(2):141-150 [FREE Full text] [doi: 10.1197/jamia.M1356] [Medline: 14633933]
8. Katz R, May L, Baker J, Test E. Redefining syndromic surveillance. J Epidemiol Glob Health 2011 Dec;1(1):21-31 [FREE Full text] [doi: 10.1016/j.jegh.2011.06.003] [Medline: 23856373]
9. Crabb BT, Lyons A, Bale M, Martin V, Berger B, Mann S, et al. Comparison of international classification of diseases and related health problems, tenth revision codes with electronic medical records among patients with symptoms of coronavirus disease 2019. JAMA Netw Open 2020 Aug 03;3(8):e2017703 [FREE Full text] [doi: 10.1001/jamanetworkopen.2020.17703] [Medline: 32797176]
10. Wagner T, Shweta F, Murugadoss K, Awasthi S, Venkatakrishnan AJ, Bade S, et al. Augmented curation of clinical notes from a massive EHR system reveals symptoms of impending COVID-19 diagnosis. Elife 2020 Jul 07;9:1-12 [FREE Full text] [doi: 10.7554/eLife.58227] [Medline: 32633720]
11. Chapman WW, Nadkarni PM, Hirschman L, D'Avolio LW, Savova GK, Uzuner O. Overcoming barriers to NLP for clinical text: the role of shared tasks and the need for additional creative solutions. J Am Med Inform Assoc 2011 Oct;18(5):540-543 [FREE Full text] [doi: 10.1136/amiajnl-2011-000465] [Medline: 21846785]
12. Friedman C, Rindflesch TC, Corn M. Natural language processing: state of the art and prospects for significant progress, a workshop sponsored by the National Library of Medicine. J Biomed Inform 2013 Oct;46(5):765-773 [FREE Full text] [doi: 10.1016/j.jbi.2013.06.004] [Medline: 23810857]
13. Velupillai S, Suominen H, Liakata M, Roberts A, Shah AD, Morley K, et al. Using clinical Natural Language Processing for health outcomes research: overview and actionable suggestions for future advances. J Biomed Inform 2018 Dec;88:11-19 [FREE Full text] [doi: 10.1016/j.jbi.2018.10.005] [Medline: 30368002]
14. Patel R, Tanwani S. Application of machine learning techniques in clinical information extraction. In: Smart Techniques for a Smarter Planet. Cham: Springer; 2019:145-165.
15. Spasic I, Nenadic G. Clinical text data in machine learning: systematic review. JMIR Med Inform 2020 Mar 31;8(3):e17984 [FREE Full text] [doi: 10.2196/17984] [Medline: 32229465]

16. Xiao C, Choi E, Sun J. Opportunities and challenges in developing deep learning models using electronic health records data: a systematic review. J Am Med Inform Assoc 2018 Oct 01;25(10):1419-1428 [FREE Full text] [doi: 10.1093/jamia/ocy068] [Medline: 29893864]

17. Mintz M, Bills S, Snow R, Jurafsky D. Distant supervision for relation extraction without labeled data. ACL AFNLP 2009. [doi: 10.3115/1690219.1690287]

18. Naseem U, Khushi M, Khan SK, Shaukat K, Moni MA. A comparative analysis of active learning for biomedical text mining. Appl Syst Innov 2021 Mar 15;4(1):23. [doi: 10.3390/asi4010023]

19. Kholghi M, Sitbon L, Zuccon G, Nguyen A. Active learning: a step towards automating medical concept extraction. J Am Med Inform Assoc 2016 Mar;23(2):289-296 [FREE Full text] [doi: 10.1093/jamia/ocv069] [Medline: 26253132]

20. Chen Y, Lasko TA, Mei Q, Denny JC, Xu H. A study of active learning methods for named entity recognition in clinical text. J Biomed Inform 2015 Dec;58:11-18 [FREE Full text] [doi: 10.1016/j.jbi.2015.09.010] [Medline: 26385377]

21. Ratner A, De SC, Wu S, Selsam D, Ré C. Data programming: creating large training sets, quickly. In: Proceedings of the 30th International Conference on Neural Information Processing Systems. 2016 Presented at: 30th International Conference on Neural Information Processing Systems; December 5 - 10, 2016; Barcelona Spain p. 3574-3582 URL: https://dl.acm.org/doi/10.5555/3157382.3157497

22. Banerjee I, Li K, Seneviratne M, Ferrari M, Seto T, Brooks JD, et al. Weakly supervised natural language processing for assessing patient-centered outcome following prostate cancer treatment. JAMIA Open 2019 Apr;2(1):150-159 [FREE Full text] [doi: 10.1093/jamiaopen/ooy057] [Medline: 31032481]

23. Fries JA, Varma P, Chen VS, Xiao K, Tejeda H, Saha P, et al. Weakly supervised classification of aortic valve malformations using unlabeled cardiac MRI sequences. Nat Commun 2019 Jul 15;10(1):3111 [FREE Full text] [doi: 10.1038/s41467-019-11012-3] [Medline: 31308376]

24. Dunnmon JA, Ratner AJ, Saab K, Khandwala N, Markert M, Sagreiya H, et al. Cross-modal data programming enables rapid medical machine learning. Patterns (N Y) 2020 May 08;1(2):100019 [FREE Full text] [doi: 10.1016/j.patter.2020.100019] [Medline: 32776018]

25. Fries J, Wu S, Ratner A, Ré C. SwellShark: a generative model for biomedical named entity recognition without labeled data. arXiv 2017 [FREE Full text]

26. Lygrisse KA, Roof MA, Keitel LN, Callaghan JJ, Schwarzkopf R, Bedard NA. The inaccuracy of ICD-10 coding in revision total hip arthroplasty and its implication on revision data. J Arthroplasty 2020 Oct;35(10):2960-2965. [doi: 10.1016/j.arth.2020.05.013] [Medline: 32507451]

27. Logan R, Davey P, De Souza N, Baird D, Guthrie B, Bell S. Assessing the accuracy of ICD-10 coding for measuring rates of and mortality from acute kidney injury and the impact of electronic alerts: an observational cohort study. Clin Kidney J 2020 Dec;13(6):1083-1090 [FREE Full text] [doi: 10.1093/ckj/sfz117] [Medline: 33391753]

28. McIsaac DI, Hamilton GM, Abdulla K, Lavallée LT, Moloo H, Pysyk C, et al. Validation of new ICD-10-based patient safety indicators for identification of in-hospital complications in surgical patients: a study of diagnostic accuracy. BMJ Qual Saf 2020 Mar;29(3):209-216. [doi: 10.1136/bmjqs-2018-008852] [Medline: 31439760]

29. Samannodi M, Hansen M, Hasbun R. Lack of accuracy of the international classification of disease, ninth (ICD-9) codes in identifying patients with encephalitis. J Neurol 2019 Apr;266(4):1034-1035. [doi: 10.1007/s00415-019-09229-9] [Medline: 30729315]

30. Horsky J, Drucker E, Ramelson H. Accuracy and completeness of clinical coding using ICD-10 for ambulatory visits. AMIA Annu Symp Proc 2017;2017:912-920 [FREE Full text] [Medline: 29854158]

31. Weng W, Wagholikar KB, McCray AT, Szolovits P, Chueh HC. Medical subdomain classification of clinical notes using a machine learning-based natural language processing approach. BMC Med Inform Decis Mak 2017 Dec 01;17(1):155 [FREE Full text] [doi: 10.1186/s12911-017-0556-8] [Medline: 29191207]

32. Xu K, Lam M, Pang J, Gao X, Band C, Mathur P, et al. Multimodal machine learning for automated ICD coding. arXiv 2018 [FREE Full text]

33. Haendel MA, Chute CG, Robinson PN. Classification, ontology, and precision medicine. N Engl J Med 2018 Dec 11;379(15):1452-1462. [doi: 10.1056/NEJMra1615014] [Medline: 30304648]

34. Stanfill MH, Williams M, Fenton SH, Jenders RA, Hersh WR. A systematic literature review of automated clinical coding and classification systems. J Am Med Inform Assoc 2010;17(6):646-651 [FREE Full text] [doi: 10.1136/jamia.2009.001024] [Medline: 20962126]

35. Baumel T, Nassour-Kassis J, Cohen R, Elhadad M, Elhadad N. Multi-label classification of patient notes a case study on ICD code assignment. arXiv 2017 [FREE Full text]

36. Shi H, Xie P, Hu Z, Zhang M, Xing E. Towards automated ICD coding using deep learning. arXiv 2017 [FREE Full text]

37. Huang J, Osorio C, Sy LW. An empirical evaluation of deep learning for ICD-9 code assignment using MIMIC-III clinical notes. Comput Methods Programs Biomed 2019 Aug;177:141-153. [doi: 10.1016/j.cmpb.2019.05.024] [Medline: 31319942]

38. Campbell S, Giadresco K. Computer-assisted clinical coding: a narrative review of the literature on its benefits, limitations, implementation and impact on clinical coding professionals. Health Inf Manag 2020 Jan;49(1):5-18. [doi: 10.1177/1833358319851305] [Medline: 31159578]

39. Goldstein I, Arzrumtsyan A, Uzuner O. Three approaches to automatic assignment of ICD-9-CM codes to radiology reports. AMIA Annu Symp Proc 2007 Oct 11:279-283 [FREE Full text] [Medline: 18693842]

40. Shickel B, Tighe PJ, Bihorac A, Rashidi P. Deep EHR: a survey of recent advances in deep learning techniques for electronic health record (EHR) analysis. IEEE J Biomed Health Inform 2018 Dec;22(5):1589-1604. [doi: 10.1109/JBHI.2017.2767063] [Medline: 29989977]

41. Xie P, Xing E. A neural architecture for automated ICD coding. In: Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics. 2018 Presented at: 56th Annual Meeting of the Association for Computational Linguistics; July, 2018; Melbourne, Australia p. 1066-1076. [doi: 10.18653/v1/p18-1098]

42. Ratner A, Bach S, Ehrenberg H, Fries J, Wu S, Ré C. Snorkel: rapid training data creation with weak supervision. VLDB J 2020;29(2):709-730. [doi: 10.1007/s00778-019-00552-1] [Medline: 32214778]

43. Honnibal M, Montani I, Van LS, Boyd A. Industrial-strength Natural Language Processing in Python. spaCy. 2020. URL: https://spacy.io/ [accessed 2022-02-03]

44. Mikolov T, Sutskever I, Chen K, Corrado G, Dean J. Distributed representations of words and phrases and their compositionality. Adv Neural Inf Process Syst 2013;3111:3119. [doi: 10.18653/v1/d16-1146]

45. Zhang Y, Chen Q, Yang Z, Lin H, Lu Z. BioWordVec, improving biomedical word embeddings with subword information and MeSH. Sci Data 2019 May 10;6(1):52 [FREE Full text] [doi: 10.1038/s41597-019-0055-0] [Medline: 31076572]

46. Le Q, Mikolov T. Distributed representations of sentences and documents. In: Proceedings of the 31st International Conference on Machine Learning. 2014 Presented at: 31st International Conference on Machine Learning; June 21–26, 2014; Beijing, China p. 1188-1196 URL: https://proceedings.mlr.press/v32/le14.html

47. Tamang S. CLEVER base terrminology. GitHub. URL: https://github.com/stamang/CLEVER [accessed 2022-02-03]

48. Neumann M, King D, Beltagy I, Ammar W. ScispaCy: fast and robust models for biomedical natural language processing. In: Proceedings of the 18th BioNLP Workshop and Shared Task. 2019 Presented at: 18th BioNLP Workshop and Shared Task; August, 2019; Florence, Italy. [doi: 10.18653/v1/w19-5034]

49. Wei C, Peng Y, Leaman R, Davis A, Mattingly C, Li J, et al. Overview of the BioCreative V Chemical Disease Relation (CDR) Task. In: Proceedings of the Fifth BioCreative Challenge Evaluation Workshop. 2015 Presented at: Fifth BioCreative Challenge Evaluation Workshop; 2015; Spain p. 154-166 URL: https://biocreative.bioinformatics.udel.edu/media/store/files/2015/BC5CDR_overview.final.pdf

50. Breiman L. Random forests. Statistics Department, University of California, Berkeley, CA. 2001. URL: https://www.stat.berkeley.edu/~breiman/randomforest2001.pdf [accessed 2022-02-10]

51. Hastie T, Friedman J, Tibshirani R. The elements of statistical learning: mining, inference, and prediction. In: Springer Series in Statistics. New York: Springer; 2001.

52. Steinkamp JM, Bala W, Sharma A, Kantrowitz JJ. Task definition, annotated dataset, and supervised natural language processing models for symptom extraction from unstructured clinical notes. J Biomed Inform 2020 Feb;102:103354 [FREE Full text] [doi: 10.1016/j.jbi.2019.103354] [Medline: 31838210]

53. Uzuner O, Solti I, Xia F, Cadag E. Community annotation experiment for ground truth generation for the i2b2 medication challenge. J Am Med Inform Assoc 2010;17(5):519-523 [FREE Full text] [doi: 10.1136/jamia.2010.004200] [Medline: 20819855]

54. Moons E, Khanna A, Akkasi A, Moens M. A comparison of deep learning methods for ICD coding of clinical records. Appl Sci 2020 Jul 30;10(15):5262. [doi: 10.3390/app10155262]

55. Johnson AE, Pollard TJ, Shen L, Lehman LH, Feng M, Ghassemi M, et al. MIMIC-III, a freely accessible critical care database. Sci Data 2016;3:160035 [FREE Full text] [doi: 10.1038/sdata.2016.35] [Medline: 27219127]

56. Miranda-Excalada A, Gonzalez-Agirre A, Armengol-Estapé J, Krallinger M. Overview of automatic clinical coding: annotations, guidelines, and solutions for non-English clinical cases at CodiEsp track of eHealth CLEF 2020. CLEF (Working Notes) 2020. 2020. URL: https://scholar.google.com/citations?view_op=view_citation&hl=en&user=1UFCgX0AAAAJ&citation_for_view=1UFCgX0AAAAJ:wbdj-CoPYUoC [accessed 2022-02-03]

57. Alsentzer E, Murphy J, Boag W, Weng W, Jin D, Naumann T, et al. Publicly available clinical BERT embeddings. In: Proceedings of the 2nd Clinical Natural Language Processing Workshop. 2019 Presented at: 2nd Clinical Natural Language Processing Workshop; June, 2019; Minneapolis, Minnesota, USA p. 72-78. [doi: 10.18653/v1/w19-1909]

58. Lee J, Yoon W, Kim S, Kim D, Kim S, So CH, et al. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. Bioinformatics 2020 Feb 15;36(4):1234-1240. [doi: 10.1093/bioinformatics/btz682] [Medline: 31501885]

59. Peng Y, Yan S, Lu Z. Transfer learning in biomedical natural language processing: an evaluation of BERT and ELMo on ten benchmarking datasets. In: Proceedings of the 18th BioNLP Workshop and Shared Task. 2019 Presented at: 18th BioNLP Workshop and Shared Task; August, 2019; Florence, Italy p. 58-65. [doi: 10.18653/v1/w19-5006]

60. Gao S, Alawad M, Young MT, Gounley J, Schaefferkoetter N, Yoon H, et al. Limitations of transformers on clinical text classification. IEEE J Biomed Health Inform 2021 Sep;25(9):3596-3607. [doi: 10.1109/jbhi.2021.3062322]

Abbreviations

AUROC: area under the receiver operating characteristic curve
BERT: bidirectional encoder representations from transformers
BOW: bag-of-words
CBOW: continuous bag-of-words
EHR: electronic health record
ICD-10: International Classification of Disease–10th revision
i2b2: Informatics for Integrating Biology and the Bedside
MIMIC: Medical Information Mart for Intensive Care
NLP: natural language processing
PV-DBOW: paragraph vector–distributed bag-of-words
SHC: Stanford Health Care
TF-IDF: term frequency–inverse document frequency

XSL•FO
RenderX

<u>Original Paper</u>

Electronic Health Record–Triggered Research Infrastructure Combining Real-world Electronic Health Record Data and Patient-Reported Outcomes to Detect Benefits, Risks, and Impact of Medication: Development Study

Karin Hek[1], PhD; Leàn Rolfes[2], PharmD, PhD; Eugène P van Puijenbroek[2,3], MD, Prof Dr; Linda E Flinterman[1], PhD; Saskia Vorstenbosch[2], MSc; Liset van Dijk[1,3], Prof Dr, IR; Robert A Verheij[1,4], Prof Dr

[1]Nivel, Netherlands Institute for Health Services Research, Utrecht, Netherlands

[2]Netherlands Pharmacovigilance Centre Lareb, 's-Hertogenbosch, Netherlands

[3]Groningen Research Institute of Pharmacy, Unit of PharmacoTherapy, - Epidemiology & -Economics, University of Groningen, Groningen, Netherlands

[4]Tilburg School of Social and Behavioral Sciences (Tranzo), Tilburg University, Tilburg, Netherlands

Corresponding Author:
Karin Hek, PhD
Nivel, Netherlands Institute for Health Services Research
PO Box 1568
Utrecht, 3500 BN
Netherlands
Phone: 31 302729700
Email: k.hek@nivel.nl

Abstract

Background: Real-world data from electronic health records (EHRs) represent a wealth of information for studying the benefits and risks of medical treatment. However, they are limited in scope and should be complemented by information from the patient perspective.

Objective: The aim of this study is to develop an innovative research infrastructure that combines information from EHRs with patient experiences reported in questionnaires to monitor the risks and benefits of medical treatment.

Methods: We focused on the treatment of overactive bladder (OAB) in general practice as a use case. To develop the Benefit, Risk, and Impact of Medication Monitor (BRIMM) infrastructure, we first performed a requirement analysis. BRIMM's starting point is routinely recorded general practice EHR data that are sent to the Dutch Nivel Primary Care Database weekly. Patients with OAB were flagged weekly on the basis of diagnoses and prescriptions. They were invited subsequently for participation by their general practitioner (GP), via a trusted third party. Patients received a series of questionnaires on disease status, pharmacological and nonpharmacological treatments, adverse drug reactions, drug adherence, and quality of life. The questionnaires and a dedicated feedback portal were developed in collaboration with a patient association for pelvic-related diseases, *Bekkenbodem4All*. Participating patients and GPs received feedback. An expert meeting was organized to assess the strengths, weaknesses, opportunities, and threats of the new research infrastructure.

Results: The BRIMM infrastructure was developed and implemented. In the Nivel Primary Care Database, 2933 patients with OAB from 27 general practices were flagged. GPs selected 1636 (55.78%) patients who were eligible for the study, of whom 295 (18.0% of eligible patients) completed the first questionnaire. A total of 288 (97.6%) patients consented to the linkage of their questionnaire data with their EHR data. According to experts, the strengths of the infrastructure were the linkage of patient-reported outcomes with EHR data, comparison of pharmacological and nonpharmacological treatments, flexibility of the infrastructure, and low registration burden for GPs. Methodological weaknesses, such as susceptibility to bias, patient selection, and low participation rates among GPs and patients, were seen as weaknesses and threats. Opportunities represent usefulness for policy makers and health professionals, conditional approval of medication, data linkage to other data sources, and feedback to patients.

Conclusions: The BRIMM research infrastructure has the potential to assess the benefits and safety of (medical) treatment in real-life situations using a unique combination of EHRs and patient-reported outcomes. As patient involvement is an important

XSL•FO
RenderX

aspect of the treatment process, generating knowledge from clinical and patient perspectives is valuable for health care providers, patients, and policy makers. The developed methodology can easily be applied to other treatments and health problems.

Introduction

Background

Electronic health records (EHRs) are increasingly used for the postmarketing surveillance of medicines, including information on prescription data, health care use, and morbidity [1,2]. However, EHRs lack information on personal significance or patients' perspectives toward the use of medicines, including experienced adverse drug reactions (ADRs), and on patients' health-related quality of life. Postmarketing surveillance should provide information about clinical and patient-reported outcomes (PROs). This allows insight into the benefit-risk balance of various treatments, including medication [3].

PROs provide in-depth insights into experiences and safety issues from the patients' perspective. Such information is not routinely obtained in the premarketing phase of a medicine or during standard care. However, it would provide insight into how patients deal with their disease and treatment, and it can help patients and health care professionals in shared and informed decision-making. Thus, an infrastructure that combines clinical information from routine EHRs with patients' experiences collected through questionnaires would allow the rapid generation of real-world information about the benefit-risk balance of various treatments, including medication, considering the perspectives of patients.

The infrastructure should be simple and reliable for patients and must have a limited administrative burden on the participating health care provider. Furthermore, it should be developed in such a way that it can be easily implemented for other diseases and treatments. Primary care is a suitable setting to develop such an infrastructure because most medications are prescribed in primary care. Moreover, in countries where primary care has a gatekeeper function, patients' primary care EHR holds a complete record of morbidity and medication of a defined list of patients (most gatekeeping systems are also list systems, where general practitioner [GP] practices have a defined list of patients that they are supposed to care for), and thus, it provides an excellent opportunity to assess the benefit-risk balance of medication when complemented with PROs.

Objective

This paper describes the requirements and the development of such a research infrastructure for the Dutch primary care setting called the Benefit, Risk, and Impact of Medication Monitor (BRIMM). We reflect on developing and using the BRIMM infrastructure as well as on its strengths, weaknesses, opportunities, and threats.

We chose the overactive bladder (OAB) as the use case to set up the infrastructure and assess its feasibility. OAB is a symptom-defined condition characterized by urinary urgency, usually with increased urinary frequency, waking up during the night to urinate, and sometimes with urgency incontinence. The high prevalence of OAB [3]; the increasing number of older adults; the negative effects of OAB on health-related quality of life [4]; and the presence of a relatively new medicine indicated for the treatment of OAB (mirabegron), which is under additional monitoring by regulatory authorities [5,6], makes OAB a suitable use case to develop and test this innovative way of collecting data for a benefit-risk registry in primary care.

Methods

Setting

BRIMM combines data from EHRs collected from general practices participating in the Nivel Primary Care Database (Nivel-PCD [7]) and PROs collected via questionnaires sent out with the Lareb Intensive Monitoring (LIM) system [8]. Nivel-PCD collects EHR data from almost 10% of the Dutch population (approximately 500 general practices, 1.8 million population). Data were collected since 1996 (from a small number of practices). Nivel-PCD contains data on consultations, morbidity, prescriptions, referrals, and clinical outcomes such as blood pressure measurements. Morbidity was recorded according to the International Classification of Primary Care version 1 (ICPC codes) used by the Dutch GPs. Prescription data were recorded using the Anatomical Therapeutic Chemical (ATC) classification. Nivel-PCD receives EHR data weekly from 350 practices, with more than 1 million listed people, allowing the identification of prevalent and incident OAB cases and following them over time. Data in Nivel-PCD are pseudonymized at the source (in the practices), leaving out directly identifying data such as names or addresses [9].

LIM is a tool to collect longitudinal PROs data; for example, on the occurrence of ADRs, coping, and impact on quality of life [10]. LIM was introduced in 2006 as a web-based intensive monitoring system to complement the spontaneous reporting of ADRs [8]. Patients received web-based questionnaires at different time points. The LIM system is flexible, as tailor-made questionnaires can be designed, which allows new questions to be added easily.

Requirements

On the basis of these existing infrastructures, we began developing the BRIMM research infrastructure with a formulation of requirements. Following discussions with the project team, we formulated the infrastructure requirements shown in Textbox 1.

XSL•FO
RenderX

Textbox 1. Infrastructure requirements.

Versatile and flexible

- Not limited to one drug or treatment

Unobtrusive

- Little or no interference with usual workflow
- Little extra work for general practitioners (GPs) or practice personnel
- Low threshold for patients to participate
- On the basis of existing infrastructures if possible

Timely

- Real-time or near real-time data collection
- Can be easily changed to monitor other treatments or diseases

Useful

- Contribute to better care (quality and efficiency)
- Generates information that is useful for GPs and patients during consultations
- Feedback loops to GPs
- Feedback loops to patients

Legal

- Compliant with the current data protection and privacy standards and legislation

---

This led us to set up the BRIMM infrastructure, which is described in the following sections.

Design

Figure 1 shows a schematic overview of the workflow. GPs participating in Nivel-PCD were invited to participate in BRIMM. For participating practices, we used weekly data on prescriptions and morbidity to flag patients with prevalent and incident OAB. OAB cases were flagged on the basis of ICPC code U02 (urinary frequency/urgency) or U04 (urine incontinence) or a prescription with ATC code G04BD (drugs for urinary frequency and incontinence). Only patients aged ≥18 years were flagged, and those with prostate cancer (ICPC Y77) in their health history were excluded, as OAB is a frequent complication of prostate cancer. Prevalent cases were flagged based on information from the 3 months before study participation.

A study pseudonym was generated for each patient, which allowed for data linkage between the Nivel-PCD and LIM. Nivel-PCD holds one-way pseudonymized data only [9]; that is, Nivel cannot contact patients directly. As described elsewhere, in Nivel-PCD, it is possible via a separate extraction only accessible via a trusted third party (TTP) to link the pseudonyms with a patient identification number that is known only in the practices' domain [9]. This allowed researchers to

initially flag patients who could be eligible for the study and to let GPs subsequently decide whether they were eligible. For ineligible patients, the GP was asked to provide a reason for exclusion. GPs signed an agreement with the TTP, designating the TTP as the processor of the data. GPs provided the TTP with the eligible patient's name and address, which were required to send an invitation letter. The TTP printed the letters and sent them to the patients on behalf of the GP. The invitation contained the patient's study pseudonym.

If a patient decided to participate in BRIMM, they were enrolled in the LIM study, which aimed at collecting the PROs. The study pseudonym was entered by the patient and stored in LIM, allowing linkage between Nivel-PCD and PROs. Informed consent for study participation and data linkage was obtained during the registration process. Once registered, the patient received invitations to complete questionnaires at enrollment and every 3 months for a duration of 1 year.

This process of inviting patients was repeated monthly for participating GPs, flagging incident OAB cases. Patients' name and address at the TTP were deleted 3 months after the end of patient recruitment (January 2020). A nonresponse analysis was performed using (pseudonymized) EHR data from flagged patients. GPs received a fee for each patient enrolled in the study.

Figure 1. Benefit, Risk, and Impact of Medication Monitor workflow. T provides an estimate of the timing of the workflow. For this study, general practitioners received a monthly list of patients to check. EHR: electronic health record; GP: general practitioner; LIM: Lareb Intensive Monitoring; Nivel-PCD: Nivel Primary Care Database; OAB: overactive bladder; TTP: trusted third party.
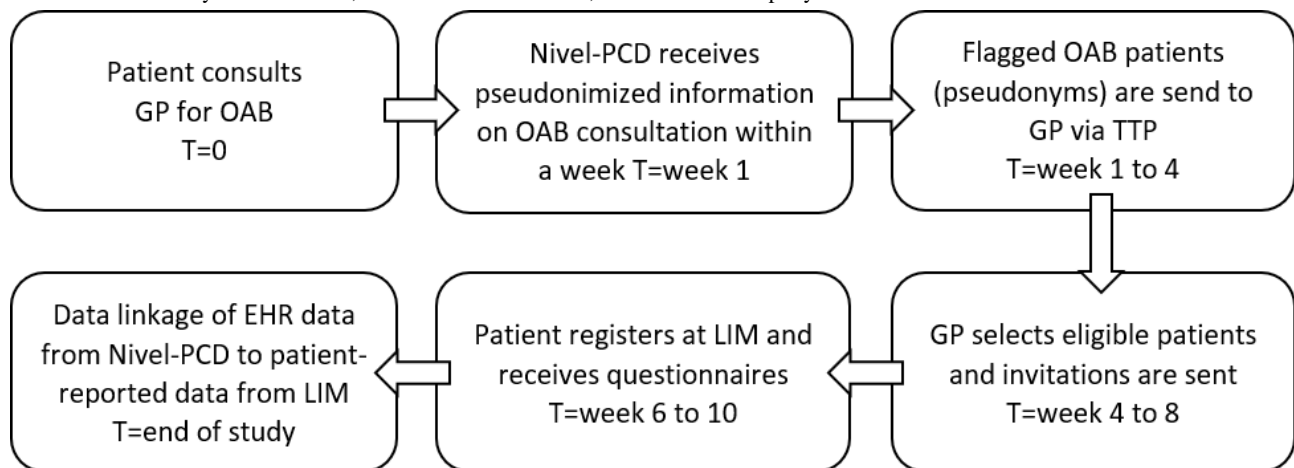
Data Collection

Patient questionnaires were developed by the project team in collaboration with an advisory board, representing key stakeholders of the patient association for pelvic-related diseases (Bekkenbodem4All), the Dutch College of General Practitioners (NHG), the Dutch Union of Urology, the Dutch Medicines Evaluation Board (MEB), a professor of pharmacy, and a health economist. Validated questionnaires were used if available. The questionnaire was only available on the web and covered the following topics. Topic A was asked in the first questionnaire, and topics B to F were included in all five questionnaires:

- Topic A: patient characteristics, including sex, year of birth, profession, education level, and socioeconomic position
- Topic B: start of OAB and contacts with health care providers for OAB
- Topic C: OAB-related symptoms (urogenital distress inventory-6 [11])
- Topic D: pharmacological and nonpharmacological treatments and treatment adherence using the Medication Adherence Rating Scale 5 [12] and the Exercise Adherence Rating Scale [13]
- Topic E: experienced ADRs
- Topic F: quality of life (EQ-5D-5L [14]) and bladder complaint–related (Incontinence Impact Questionnaire-7 [11])

Privacy and Ethics Approval

Nivel and Lareb both maintain strict privacy protocols, which are also applied for this project. The email address of patients, which was required to send questionnaires, was provided by the patients upon participation in BRIMM. This information was encrypted and stored separately from the information collected in the questionnaires and EHRs. Nivel-PCD does not contain any patient-identifying information. Data were handled according to the General Data Protection Regulation to protect the privacy of participating patients and practices. The results cannot be traced back to individual persons, health care providers, or health care organizations.

This study has been approved according to the governance code of the Nivel-PCD (NZR-00316.050). The workflow was approved by the privacy committee of Nivel-PCD. The study protocol was assessed by the Medical Ethical Committee of Amsterdam Medical Center, location VUmc, who confirmed that the Medical Research Involving Human Subjects Act (in Dutch: Wet medisch-wetenschappelijk onderzoek met mensen [WMO]) does not apply to this study (#2017.506).

Governance and Data Sharing

The governance structure of Nivel-PCD applies to BRIMM, which includes a steering committee, a privacy committee, and so-called chambers with representatives of health care providers. These chambers determine the use of data. For this study, an advisory board was set up, the composition of which has been described earlier. The advisory board participated actively in the project. In addition, we set up a patient panel to advise on the content of the questionnaires and patient feedback. The linked data resulting from this registry are available for use by parties other than Lareb and Nivel after approval by the governance structure of Nivel-PCD and Lareb.

Feedback for Patients and Practices

For patients, two types of feedback were made available on their personal webpage where they filled in the questionnaires: (1) a PDF file of their filled questionnaire, which they could discuss with their health care provider, and (2) a graphic representation of some questions over time. The latter provided insights into changes in the patient's OAB status, quality of life over time, and ADRs and experiences. Feedback in the graphic presentation was updated automatically after the questionnaire had been completed. The content was determined in collaboration with the patient association Bekkenbodem4All, and 10 patients provided feedback. Feedback to the patients was provided in (near) real time.

GPs received feedback on participating patients with OAB in their practice in comparison with other practices. Feedback for GPs provided insight into the number of participating patients, their prescribing habits, and the reported ADRs. Feedback to GPs was provided at the end of the study period. The feedback

topics included health care utilization, adverse effects, adherence to treatment, and quality of life.

Pilot Study

A pilot study was performed in 2 GP practices between January 2018 and August 2018 to test the feasibility of the infrastructure and to evaluate whether BRIMM could be improved from the perspective of GPs. Both practices were asked about their experiences with BRIMM in a semistructured telephone interview with the GP (practice 1) and practice nurse (practice 2).

OAB Use Case

The pilot study was followed by a use case study in which GPs from the Nivel-PCD network were recruited. The GPs were invited by email and in groups of 20-30 practices. The invitation material was updated several times to test which material worked best. We emailed a flyer with a short description of the study, a link to a presentation with highlights, and an elaborate description of the study. Participating GPs received a frequently asked question document for practice personnel and a movie to be played in the GPs' waiting room to introduce the study to their patients. A total of 27 GPs (including the 2 pilot practices) participated in the study. Patient nonresponse was described using descriptive statistics.

Evaluative Strengths, Weaknesses, Opportunities, and Threats Analysis

In September 2018, Nivel and Lareb hosted a stakeholders meeting to inform stakeholders about BRIMM and retrieve their views on BRIMM. A total of 10 stakeholders attended the meeting representing a variety of institutions, including professional associations for GPs and pharmacists, the patient federation, pharmaceutical companies, pharmacovigilance center Lareb, the Medicines Evaluation Board, the national institutes on rational medicine use, and public health and environment. The BRIMM infrastructure and the results of the pilot study were presented at the meeting, after which 2 groups of stakeholders were formed and asked to evaluate the strengths, weaknesses, opportunities, and threats of BRIMM. The results of this evaluation are presented herein.

Results

Pilot Study

The patient recruitment route was successfully tested in the 2 practices that participated in the pilot study. A GP and a practice nurse checked their list of flagged patients monthly and provided the addresses of eligible patients. Both the GP and the practice nurse reported that the required time was as expected and that this was feasible in practice. They were also content with the information beforehand and the way and speed with which minor technical issues were handled, and they had no suggestions to further improve the patient recruitment route. They appreciated that the patients were referred to Nivel and Lareb for questions. Both practices would advise colleagues to participate in this project. The pilot study did not lead to any changes in the study design.

OAB Use Case Results

A total of 27 GP practices participated in the study. In Nivel-PCD, 2933 patients with OAB were flagged. GPs selected 1636 patients (55.78%) who were eligible for the study, and 358 (21.88% of the eligible patients) registered for the study, of whom 295 (18.03% of the eligible patients) completed the first questionnaire. Practices recruited between 0 and 40 patients, with a mean of 10 (median 6, IQR 5-17) patients per practice. A total of 7 patients did not provide consent for data linkage; therefore, results that were calculated using EHR data were based on the 288 patients who provided consent for data linkage.

The main reason for excluding patients was no diagnosis of OAB (539/1297, 41.6% of exclusions; Table 1). Participating patients were, on average, slightly older (mean age 66, SD 13 years) and less often females (141/295, 47.8%) than those who were flagged (mean age 63, SD 19 years; 1741/2933, 59.36% females) and those who were invited (mean age 63, SD 18 years; 958/1636, 58.56% females; Table 2). The use of OAB medications was higher among participants, and participants had fewer chronic conditions than patients who were invited or flagged (Table 2).

We collected information on health care use, bladder complaints, and ADRs from 295 participants. The second questionnaire was completed by 163 (55.2%) patients. Of these, 102 (34.6%) patients completed all 5 questionnaires.

Table 1. Reasons for general practitioners to exclude flagged patients (n=1297).

Exclusion criteria	Patients, n (%)
No overactive bladder	539 (41.56)
Reason unknown	315 (24.29)
Cognitively or mentally unable	142 (10.95)
Moved or deceased	82 (6.32)
Cannot handle a personal computer	63 (4.86)
Treated by a urologist	53 (4.09)
Terminally ill or in hospital	46 (3.55)
Insufficient knowledge of the Dutch language	37 (2.85)
Other reasons	20 (1.54)

XSL•FO
RenderX

Table 2. Characteristics of the study population.

Characteristics	Flagged patients (n=2933)	Invited patients (n=1636)	Participating patients (n=295)[a]
Female, n (%)	1741 (59.36)	958 (58.56)	141 (47.77)
Age (years), mean (SD)	63.46 (19.05)	62.91 (17.89)	66.4 (12.91)
Age (years), n (%)			
18-44	549 (18.72)	282 (17.24)	19 (6.42)
45-64	758 (25.84)	457 (27.93)	95 (32.32)
65-74	652 (22.23)	423 (25.86)	111 (37.58)
75-84	620 (21.14)	334 (20.42)	54 (18.32)
≥85	354 (12.07)	140 (8.56)	16 (5.43)
Use of overactive bladder medication, n (%)[b,c]	640 (22.42)	392 (24.87)	68 (25.78)
Number of chronic diseases, n (%)[b,d]			
0	405 (20.94)	203 (20.34)	31 (16.93)
1 or 2	725 (27.48)	421 (42.18)	80 (43.69)
≥3	804 (41.57)	374 (37.47)	72 (39.31)

[a]Results of the 7 patients who did not give consent for data linkage between questionnaires and electronic health record data were not included in calculations on overactive bladder medication and number of chronic diseases.

[b]Information on medication use was available for 2854, 1576, and 264 patients, and information on chronic comorbidities was available for 1934, 998, and 183 patients.

[c]Overactive bladder medication was defined as medication with Anatomical Therapeutic Chemical code G04BD.

[d]The number of chronic comorbidities was based on a list of the 29 most common chronic diseases, including, for example, Chronic Obstructive Pulmonary Disease and cardiovascular disease [15].

Results of Strengths, Weaknesses, Opportunities, and Threats Analysis

Figure 2 provides a summary of the main findings of the Strengths, Weaknesses, Opportunities, and Threats analysis based on a discussion with stakeholders.

Figure 2. Summary of strengths, weaknesses, opportunities, and threats of the Benefit, Risk, and Impact of Medication Monitor infrastructure. BRIMM: Benefit, Risk, and Impact of Medication Monitor; GP: general practitioner.

Strengths

- Enrichment of data with patient experiences
- Comparison of treatments, including nonpharmacological treatment
- More knowledge on medication interaction
- A structured questionnaire can give patients new insights
- Heterogeneous population (generalizability)
- Flexible system
- Detailed nonresponse analysis possible
- No additional registration burden for GPs

Weaknesses

- Following patients over time can also give misleading information
- The association between outcome and treatment is difficult to infer
- Requires digital skills from patients
- Selection of GP practices and patients (participation rate)
- Practice and patient recruitment takes time
- Susceptible to bias (internal validity)

BRIMM

Opportunities

- Provide information for policy makers and health system
- Data linkage to other real-world data sources
- Feedback to patients
- Data for conditional approval of medication
- Patient are more honest in questionnaires than in interviews, feedback to GP possible
- Link to individual care plan

Threats

- Data linkage may feel like *big brother* for patients
- Selection of patients
- Data do not necessarily reflect knowledge and actionable information

Strengths

A unique selling point of BRIMM is the combination of EHR data with patient experiences to compare pharmacological and nonpharmacological treatments. These data can be used to explore clinical research questions concerning specific diseases or treatments, patient perspectives, and economic evaluations. By including EHR data, BRIMM provides reliable information about concomitant medications and comorbidities. Moreover, by making efficient use of the EHR data during patient preselection, the final patient selection by the GP takes only a limited amount of time. Having EHR data for all patients in the preselection also enables a detailed nonresponse analysis and possible selection bias. Furthermore, the flexibility of BRIMM is an asset, and it can be adapted to other diseases or medications. Finally, a strength is the feedback that patients receive from their structured questionnaires. These may be valuable to patients, as it may provide a better understanding of their disease and the impact of treatment and support communication with the GP.

Weaknesses

A weakness that potentially affects study outcomes is that the association between outcomes and treatment is difficult to infer. Furthermore, this type of study is prone to selection bias for instance because of selective nonresponse by both GPs and patients. Both practice and patient recruitment requires time. This makes BRIMM less appropriate for answering acute, urgent research questions.

Opportunities

BRIMM may provide important information for policy makers and the health system and for individual (participating) patients and GPs in the form of feedback information and study outcomes. The data collected can be linked to other real-world data sources, and the outcomes of the study can be used to design individual patient care plans. In the future, BRIMM could be used for dedicated studies of drugs with conditional approval.

Threats

The stakeholders identified 3 possible threats. First, patients may distrust data linkage of EHR data with their questionnaire data, as they may feel that they are not in control of the nature of exchange of the data. Second, patient selection may affect the validity of the study outcomes. Finally, the amount of data collected does not automatically reflect the knowledge or actionable information.

Discussion

Overview

BRIMM is a unique research infrastructure that combines routinely recorded primary care EHRs with PROs to monitor the safety and benefits of medical treatment. Till date, studies have generally focused only on EHR data or PROs. By combining both types of data, we are able to provide a more complete picture of the patient characteristics and disease status (clinically and from a patient perspective), benefits and risks of treatments, drug use adherence, and quality of life. It also allows for a long-term follow-up of patients, which makes it possible to explore the long-term benefits and risks of treatments. BRIMM fits in with the trend of valorizing real-world databases [16,17].

BRIMM met our predefined criteria that it should be easy to use, timely, flexible, useful, and in accordance with legal requirements; however, we also encountered some practical challenges. Here, we reflect on the lessons learned from developing and testing the BRIMM research infrastructure that can be implemented for future studies combining EHR and PRO data.

Lessons Learned

We extensively explored requirements for the BRIMM research infrastructure and composed general requirements to ensure the infrastructure would be widely usable, not only for the OAB use case. This means that this infrastructure can now be relatively easily implemented for other studies. For example, it took only a few months to prepare a second study on COVID-19 using the BRIMM method. For some studies, this might not be quick enough. Preparations, including development of questionnaire and ethical approval procedure, and GP and patient inclusion will take at least several months. Including patients for the BRIMM infrastructure only takes limited time for GPs. However, for any future studies, it should be considered that GPs have many other activities, and workload in general is an important barrier for participating in studies [18]. Thus, we pursued the GPs to participate by giving feedback, a fee, and highlighting the relevance of BRIMM in different forms (presentation, movie, and infographic). These efforts led to 27 GPs being included in the OAB use case. The study topic may also play an important role in the willingness of GPs to participate. GPs told us that OAB was not a topic of societal interest. On the contrary, an invitation for participation in a BRIMM study focused on patients with COVID-19 received interest of almost 100 general practices. Therefore, we recommend conducting an inventory among GP practices to attain their interest in the subject before the start of a next BRIMM-like study.

Stakeholders identified patients' reluctance to consent to data linkage as a possible threat. However, this was not observed in this study, as 97.6% (288/295) of the patients consented to data linkage, which is in line with findings elsewhere [19,20]. For patient selection, it should be noted that BRIMM is set off by EHR data (ie, ICPC codes and ATC codes). The granularity of coding systems has been an issue in OAB use cases. The specificity of ICPC codes to include patients was limited; that is, there was no specific ICPC code for OAB. Therefore, GPs excluded a large group of initially flagged patients who did not appear to have OAB. In addition, we initially flagged some patients who should have been excluded because of prostate cancer. Apparently, prostate cancer had not been recorded in the EHR. This makes the infrastructure more suitable for diagnoses that can be flagged using a specific ICPC code or an ATC code. The more precise the initial flagging of patients, the less time consuming the efforts of GPs to check whether patients are indeed eligible.

Furthermore, the response rate of the patients was relatively low (295/1636, 18.03%). In general, response rates have declined over the past few years. Response rates may be increased by providing an incentive for patients to participate [21] and by using paper instead of web-based questionnaires [22,23]. However, web-based PRO collection also has several advantages, including lower costs and higher data quality [23,24].

Patient participation during the setup of the BRIMM research infrastructure was deemed important to ensure that the study benefits this patient group. However, involving patients in this study appeared to be a challenge. It took relatively much effort to receive feedback on the questionnaire. The feedback provided to the patients was hardly used by them, limiting the benefits for participating patients that we hoped for. Therefore, before conducting a study, it should be explored whether patients would like to receive feedback at all, and if so, what type of feedback they would like and how it should be made available to them. In addition, one can argue that providing feedback during the study may bias patients' consecutive answers to questionnaires and, therefore, also the results. Moreover, this may be aimed for when providing actionable feedback. The infrastructure could then be regarded as an intervention as well as a research tool, along the lines of a learning health system [25-27]. However, we do not expect that the feedback in this study had biased the results, as the feedback was designed to provide a basic overview of adverse effects and bladder complaints and was hardly used. Depending on its aim, future studies should consider whether feedback should be provided during or at the end of the study.

The BRIMM research infrastructure can be used to investigate the benefits and risks of almost any treatment for which citizens consult their GP and may, for example, provide valuable information to support supplemental indications; that is, permission for medicines already on the market to be used for new indications, patient groups, or stages of disease [16,28,29]. However, to allow for the efficient use of resources, it is important to choose diseases or medicines with a high incidence, prevalence, or use in primary care to allow for the inclusion of a sufficient number of patients.

Other Applications of the BRIMM Infrastructure

BRIMM's unique combination of EHR and PRO data can also be used for other purposes than the monitoring of benefits and risks of medication, for example, to investigate the patient's perspective regarding diseases and their treatments, such as the long-term effects of COVID-19 and the long-term effects of implants, for example, breast implants, or to study the effects of over-the-counter medication and home remedies for diseases. Furthermore, the BRIMM research infrastructure could be used to efficiently collect clinical and PRO data for clinical trials. Similar innovative methods to stimulate patient recruitment have been proposed before [30-32]. Finally, we will further analyze the data collected in the OAB use case to ensure the data are converted to actionable information.

The BRIMM research infrastructure was implemented in the Netherlands, a country with a primary care–oriented health care system and widespread use of relatively well-developed EHR systems in general practice. The infrastructure can also be implemented in other countries. Minimal requirements are a well-developed EHR system, digitally literate patient population, and sufficient quality of routinely recorded health data [33]. Furthermore, it is important to consider the position of GPs in the health care system, particularly with respect to the disease or treatment studied. Implementation in a primary care–oriented gatekeeper system, for example, and focusing on a disease that is primarily treated in primary care might be preferable. Within the Netherlands, the system can be implemented in other EHR databases and combined with PRO measures using the workflow described here.

Conclusions

The BRIMM research infrastructure makes it possible to assess the benefits and safety of (medical) treatment in real-life health care situations using a unique combination of EHRs and PROs, and it does so in a nonobtrusive way, without causing much extra administrative burden for health care professionals. As patient involvement is an important aspect of the treatment process, generating knowledge from both the clinical and patient perspectives is valuable for both health care providers and patients.

BRIMM has significant methodological advantages to serve as a tool for postmarketing surveillance of drugs that require additional monitoring. In addition, it provides enhanced possibilities to create cohorts of patients to conduct large-scale comparative effectiveness research to accomplish the goals of a learning health system that supports patients, physicians, and policy makers in making informed decisions [25]. BRIMM can in principle be applied to any type of disease or medicine in primary care.

Conflicts of Interest

References

1. Charlton R, Snowball J, Sammon C, de Vries C. The clinical practice research datalink for drug safety in pregnancy research: an overview. Therapie 2014;69(1):83-89. [doi: 10.2515/therapie/2014007] [Medline: 24698192]

2. Moore N, Berdaï D, Blin P, Droz C. Pharmacovigilance - the next chapter. Therapie 2019;74(6):557-567. [doi: 10.1016/j.therap.2019.09.004] [Medline: 31623850]

3. Irwin DE, Kopp ZS, Agatep B, Milsom I, Abrams P. Worldwide prevalence estimates of lower urinary tract symptoms, overactive bladder, urinary incontinence and bladder outlet obstruction. BJU Int 2011;108(7):1132-1138. [doi: 10.1111/j.1464-410X.2010.09993.x] [Medline: 21231991]

4. Stewart WF, Van Rooyen JB, Cundiff GW, Abrams P, Herzog AR, Corey R, et al. Prevalence and burden of overactive bladder in the United States. World J Urol 2003;20(6):327-336. [doi: 10.1007/s00345-002-0301-4] [Medline: 12811491]

5. Nitti VW, Khullar V, van Kerrebroeck P, Herschorn S, Cambronero J, Angulo JC, et al. Mirabegron for the treatment of overactive bladder: a prespecified pooled efficacy analysis and pooled safety analysis of three randomised, double-blind, placebo-controlled, phase III studies. Int J Clin Pract 2013;67(7):619-632 [FREE Full text] [doi: 10.1111/ijcp.12194] [Medline: 23692526]

6. Betmiga public assessment report: mirabegron. European Medicines Agency. 2012. URL: https://www.ema.europa.eu/en/medicines/human/EPAR/betmiga [accessed 2021-12-17]

7. Nivel Primary Care Database. Nivel. 2020. URL: https://nivel.nl/en/nivel-primary-care-database [accessed 2021-12-17]

8. Härmark L, van Grootheest K. Web-based intensive monitoring: from passive to active drug surveillance. Expert Opin Drug Saf 2012;11(1):45-51. [doi: 10.1517/14740338.2012.629184] [Medline: 22007719]

9. Kuchinke W, Ohmann C, Verheij RA, van Veen EB, Arvanitis TN, Taweel A, et al. A standardised graphic method for describing data privacy frameworks in primary care research using a flexible zone model. Int J Med Inform 2014;83(12):941-957. [doi: 10.1016/j.ijmedinf.2014.08.009] [Medline: 25241154]

10. Härmark LV. Web-based intensive monitoring: a patient based pharmacovigilance tool. Groningen, The Netherlands: Sine nomine; 2012.

11. Utomo E, Korfage IJ, Wildhagen MF, Steensma AB, Bangma CH, Blok BF. Validation of the urogenital distress inventory (UDI-6) and incontinence impact questionnaire (IIQ-7) in a Dutch population. Neurourol Urodyn 2015;34(1):24-31. [doi: 10.1002/nau.22496] [Medline: 24167010]

12. Chan AH, Horne R, Hankins M, Chisari C. The medication adherence report scale: a measurement tool for eliciting patients' reports of nonadherence. Br J Clin Pharmacol 2020;86(7):1281-1288 [FREE Full text] [doi: 10.1111/bcp.14193] [Medline: 31823381]

13. Newman-Beinart NA, Norton S, Dowling D, Gavriloff D, Vari C, Weinman JA, et al. The development and initial psychometric evaluation of a measure assessing adherence to prescribed exercise: the Exercise Adherence Rating Scale (EARS). Physiotherapy 2017;103(2):180-185. [doi: 10.1016/j.physio.2016.11.001] [Medline: 27913064]

14. Janssen MF, Pickard AS, Golicki D, Gudex C, Niewada M, Scalone L, et al. Measurement properties of the EQ-5D-5L compared to the EQ-5D-3L across eight patient groups: a multi-country study. Qual Life Res 2013;22(7):1717-1727 [FREE Full text] [doi: 10.1007/s11136-012-0322-4] [Medline: 23184421]

15. Sinnige J, Korevaar JC, Westert GP, Spreeuwenberg P, Schellevis FG, Braspenning JC. Multimorbidity patterns in a primary care population aged 55 years and over. Fam Pract 2015;32(5):505-513 [FREE Full text] [doi: 10.1093/fampra/cmv037] [Medline: 26040310]

16. Baumfeld Andre E, Reynolds R, Caubel P, Azoulay L, Dreyer NA. Trial designs using real-world data: the changing landscape of the regulatory approval process. Pharmacoepidemiol Drug Saf 2020;29(10):1201-1212 [FREE Full text] [doi: 10.1002/pds.4932] [Medline: 31823482]

17. Sherman RE, Anderson SA, Dal Pan GJ, Gray GW, Gross T, Hunter NL, et al. Real-world evidence - what is it and what can it tell us? N Engl J Med 2016;375(23):2293-2297. [doi: 10.1056/NEJMsb1609216] [Medline: 27959688]

18. Rosemann T, Szecsenyi J. General practitioners' attitudes towards research in primary care: qualitative results of a cross sectional study. BMC Fam Pract 2004;5(1):31 [FREE Full text] [doi: 10.1186/1471-2296-5-31] [Medline: 15613246]

19. Coppen R, van Veen EB, Groenewegen PP, Hazes JM, de Jong JD, Kievit J, et al. Will the trilogue on the EU Data Protection Regulation recognise the importance of health research? Eur J Public Health 2015;25(5):757-758 [FREE Full text] [doi: 10.1093/eurpub/ckv149] [Medline: 26265364]

20. Hutchings E, Loomes M, Butow P, Boyle FM. A systematic literature review of attitudes towards secondary use and sharing of health administrative and clinical trial data: a focus on consent. Syst Rev 2021;10(1):132 [FREE Full text] [doi: 10.1186/s13643-021-01663-z] [Medline: 33941282]

XSL•FO

RenderX

21. Hathaway CA, Chavez MN, Kadono M, Ketcher D, Rollison DE, Siegel EM, et al. Improving electronic survey response rates among cancer center patients during the COVID-19 pandemic: mixed methods pilot study. JMIR Cancer 2021;7(3):e30265 [FREE Full text] [doi: 10.2196/30265] [Medline: 34156965]

22. Kongsved SM, Basnov M, Holm-Christensen K, Hjollund NH. Response rate and completeness of questionnaires: a randomized study of Internet versus paper-and-pencil versions. J Med Internet Res 2007;9(3):e25 [FREE Full text] [doi: 10.2196/jmir.9.3.e25] [Medline: 17942387]

23. Ebert JF, Huibers L, Christensen B, Christensen MB. Paper- or web-based questionnaire invitations as a method for data collection: cross-sectional comparative study of differences in response rate, completeness of data, and financial cost. J Med Internet Res 2018;20(1):e24 [FREE Full text] [doi: 10.2196/jmir.8353] [Medline: 29362206]

24. Meirte J, Hellemans N, Anthonissen M, Denteneer L, Maertens K, Moortgat P, et al. Benefits and disadvantages of electronic patient-reported outcome measures: systematic review. JMIR Periop Med 2020;3(1):e15588 [FREE Full text] [doi: 10.2196/15588] [Medline: 33393920]

25. Platt JE, Raj M, Wienroth M. An analysis of the learning health system in its first decade in practice: scoping review. J Med Internet Res 2020;22(3):e17026 [FREE Full text] [doi: 10.2196/17026] [Medline: 32191214]

26. Friedman CP, Wong AK, Blumenthal D. Achieving a nationwide learning health system. Sci Transl Med 2010;2(57):57cm29. [doi: 10.1126/scitranslmed.3001456] [Medline: 21068440]

27. Menear M, Blanchette MA, Demers-Payette O, Roy D. A framework for value-creating learning health systems. Health Res Policy Syst 2019;17(1):79 [FREE Full text] [doi: 10.1186/s12961-019-0477-3] [Medline: 31399114]

28. Anglemyer A, Horvath HT, Bero L. Healthcare outcomes assessed with observational study designs compared with those assessed in randomized trials. Cochrane Database Syst Rev 2014;2014(4):MR000034 [FREE Full text] [doi: 10.1002/14651858.MR000034.pub2] [Medline: 24782322]

29. Fralick M, Kesselheim AS, Avorn J, Schneeweiss S. Use of health care databases to support supplemental indications of approved medications. JAMA Intern Med 2018;178(1):55-63 [FREE Full text] [doi: 10.1001/jamainternmed.2017.3919] [Medline: 29159410]

30. Zimmerman LP, Goel S, Sathar S, Gladfelter CE, Onate A, Kane LL, et al. A novel patient recruitment strategy: patient selection directly from the community through linkage to clinical data. Appl Clin Inform 2018;9(1):114-121 [FREE Full text] [doi: 10.1055/s-0038-1625964] [Medline: 29444537]

31. James S, Rao SV, Granger CB. Registry-based randomized clinical trials--a new clinical trial paradigm. Nat Rev Cardiol 2015;12(5):312-316. [doi: 10.1038/nrcardio.2015.33] [Medline: 25781411]

32. Delvaux N, Aertgeerts B, van Bussel JC, Goderis G, Vaes B, Vermandere M. Health data for research through a nationwide privacy-proof system in Belgium: design and implementation. JMIR Med Inform 2018;6(4):e11428 [FREE Full text] [doi: 10.2196/11428] [Medline: 30455164]

33. Verheij RA, Curcin V, Delaney BC, McGilchrist MM. Possible sources of bias in primary care electronic health record data use and reuse. J Med Internet Res 2018;20(5):e185 [FREE Full text] [doi: 10.2196/jmir.9134] [Medline: 29844010]

Abbreviations

ADR: adverse drug reaction
ATC: Anatomical Therapeutic Chemical
BRIMM: Benefit, Risk, and Impact of Medication Monitor
EHR: electronic health record
GP: general practitioner
ICPC: International Classification of Primary Care
LIM: Lareb Intensive Monitoring
Nivel-PCD: Nivel Primary Care Database
OAB: overactive bladder
PRO: patient-reported outcome
TTP: trusted third party

<u>Original Paper</u>

Patient-Level Fall Risk Prediction Using the Observational Medical Outcomes Partnership's Common Data Model: Pilot Feasibility Study

Hyesil Jung[1], PhD; Sooyoung Yoo[1], PhD; Seok Kim[1], MPH; Eunjeong Heo[1], BSc; Borham Kim[1], BSN; Ho-Young Lee[1], PhD; Hee Hwang[2], PhD

[1]Office of eHealth Research and Business, Seoul National University Bundang Hospital, Seongnam-si, Republic of Korea
[2]Kakao Healthcare Company-In-Company, Seongnam-si, Republic of Korea

Corresponding Author:
Sooyoung Yoo, PhD
Office of eHealth Research and Business
Seoul National University Bundang Hospital
172 Dolma-ro, Bundang-gu
Seongnam-si, 13620
Republic of Korea
Phone: 82 31 787 8980
Email: <u>yoosoo0@snubh.org</u>

Abstract

Background: Falls in acute care settings threaten patients' safety. Researchers have been developing fall risk prediction models and exploring risk factors to provide evidence-based fall prevention practices; however, such efforts are hindered by insufficient samples, limited covariates, and a lack of standardized methodologies that aid study replication.

Objective: The objectives of this study were to (1) convert fall-related electronic health record data into the standardized Observational Medical Outcome Partnership's (OMOP) common data model format and (2) develop models that predict fall risk during 2 time periods.

Methods: As a pilot feasibility test, we converted fall-related electronic health record data (nursing notes, fall risk assessment sheet, patient acuity assessment sheet, and clinical observation sheet) into standardized OMOP common data model format using an extraction, transformation, and load process. We developed fall risk prediction models for 2 time periods (within 7 days of admission and during the entire hospital stay) using 2 algorithms (least absolute shrinkage and selection operator logistic regression and random forest).

Results: In total, 6277 nursing statements, 747,049,486 clinical observation sheet records, 1,554,775 fall risk scores, and 5,685,011 patient acuity scores were converted into OMOP common data model format. All our models (area under the receiver operating characteristic curve 0.692-0.726) performed better than the Hendrich II Fall Risk Model. Patient acuity score, fall history, age ≥60 years, movement disorder, and central nervous system agents were the most important predictors in the logistic regression models.

Conclusions: To enhance model performance further, we are currently converting all nursing records into the OMOP common data model data format, which will then be included in the models. Thus, in the near future, the performance of fall risk prediction models could be improved through the application of abundant nursing records and external validation.

Introduction

Falls are the most commonly reported accidents that threaten patient safety in hospitals, particularly, because they may result in serious injuries—hip fractures and head injuries—or even death. Additionally, injurious falls increase hospital stays by up to 6 to 12 days and medical expenditures by $19,376 to $32,315 [1]. In 2015, the United States spent approximately $50 billion in fall-related additional medical costs [2]. However, most falls are considered preventable accidents, and since inpatient fall prevention depends on nursing quantity and quality, nurses have a key role.

Nurses periodically assess the risk of falls using screening tools such as the Hendrich II Fall Risk Model [3] and Morse Fall Scale [4] and provide additional nursing interventions. Furthermore, there have been ongoing attempts to improve the predictive performance of existing fall risk screening tools or develop a new prediction model altogether. Jung et al [5] improved fall prediction by integrating electronic health record data reflecting different types of data that were recorded over time and integrated from various sources. However, the participants were patients admitted to specific departments in a single hospital for a specific short period. One study [6] incorporated longitudinal electronic medical records and nursing data as covariates in calculating the fall risk and tested the model's performance through external validation. Nevertheless, the study had a limitation, in that, it did not comprehensively consider latent factors such as clinical test results. Marier et al [7] used structured electronic medical record data and the minimum data set to predict falls in nursing homes. These existing fall risk prediction models are limited because they were developed using small samples and limited covariates. Additionally, they lack standardized methodologies that allow their results to be reproduced by other researchers.

To overcome these limitations, Reps et al [8] proposed a standardized machine learning framework to generate and evaluate a clinical prediction model that leverages standardized clinical databases. Observational Health Data Science and Informatics (OHDSI) has created and applied an open-source data format and standardized analytics solutions to a diverse range of health and medical databases worldwide. The Observational Medical Outcome Partnership's (OMOP) common data model transforms heterogeneous source data into a common format using a set of common terminologies, vocabularies, and coding schemes. Thus, the OMOP common data model allows researchers to analyze health care big data from multiple sites consistently for development and replication [8]. In 2016, electronic health record data that included long-term care minimum data, drug dispensing data, and fall incident data from 5 skilled nursing facilities were converted into the OMOP common data model format [9]. Although the onset of major depressive disorder after beta-blocker therapy [10], symptomatic hemorrhagic transformation in patients with acute ischemic stroke [11], and cardioneurometabolic disease from full-night polysomnographic tests of patients [12] have been predicted using OMOP common data model data, there has been no attempt to predict inpatient fall risk in acute care settings using OMOP common data model data.

Therefore, the objectives of this study were to (1) convert fall-related electronic health record data into the standardized OMOP common data model data format and (2) develop a model that predicts fall risk at 2 risk time periods within 7 days of admission and during entire hospital stays, using OMOP common data model data as a pilot feasibility test.

Methods

Data Source

We used OMOP common data model data from Seoul National University Bundang Hospital, a tertiary general hospital located in a South Korean metropolis. Deidentified electronic health record data for more than 2 million patients who visited the hospital from May 2003 to July 2019 (patient demographic data, visit information, diagnoses, chief complaints, medications, test orders or results, interventions or surgeries, and family or past medical histories) were converted to the OMOP common data model format (version 5.3).

The OMOP common data model format consists of tables in which events of a different nature are stored. Signs, symptoms, and diagnosis are recorded in the *CONDITION_OCCURRENCE* table; activities or processes of a diagnostic or therapeutic nature ordered, or carried out by, a health care provider are recorded in the *PROCEDURE_OCCURRENCE* table; exposure to a drug (ingested or otherwise introduced into the body) are recorded in the *DRUG_EXPOSURE* table; clinical facts (including social and lifestyle facts and medical and family history) about patients obtained in the context of examination or interview are recorded in the *OBSERVATION* table; and orders and the results of laboratory tests, vital signs, and quantitative findings from pathology reports are recorded in the *MEASUREMENT* table. Events where persons or patients visit the health care system for a duration of time (for example, inpatient, outpatient, or emergency room visits) with detailed information are recorded in the *VISIT_OCCURENCE* and *VISIT_DETAIL* tables, respectively. All tables are linked to the *PERSON* table, which includes each person or patient and some demographic information, providing a person-centric relational data model [13].

Study Population

The target population consisted of patients over 18 years admitted to neurology, neurosurgery, hematology, or oncology departments from January 1, 2010 to July 18, 2019 at Seoul National University Bundang Hospital. Accidental falls have most frequently occurred in these departments in this hospital. The outcome cohort consisted of patients who had experienced falls. Patients who had a fall were identified by using 9 structured and standardized statements. Since 2003, the Seoul National University Bundang Hospital has used standardized nursing statements, with a unique predefined code that is built on the International Classification for Nursing Practice, which have been validated in previous fall-related studies [5,14]. We also searched the free-text narratives that included the words or phrases "fall down," "slip and fall," and "collapsed" to identify patients who had falls but had not been highlighted by the 9 structured and standardized statements. To identify to which medical departments patients had been admitted, we used

the specialty data of the physician (provider) who treated the patient by combining the *VISIT_OCCURRENCE* and *PROVIDER* tables, first, or care site data, obtained from the *VISIT_DETAIL* table. If there were no care site data in the *VISIT_DETAIL* table, we used ward information from the *visit_detail_source_value* field (Multimedia Appendix 1).

Conversion of Fall-Related Electronic Health Record Data Into OMOP Common Data Model Format

Fall-related electronic health record data were converted into OMOP common data model format through an extraction, transformation, and load process. We extracted the data from the nursing notes, fall risk assessment, patient acuity assessment, and clinical observation sheets. Next, we integrated and standardized the data in the common data model format using standardized terminologies (SNOMED CT and LOINC). Nine structured nursing statement items and free-text narratives that included the words or phrases "fall down," "slip and fall," and "collapsed" in the nursing notes were manually mapped to 3 standard concepts (*SCTID = 33036003 |Fall on same level (event)|*, *SCTID = 242120009 |Fall on public service vehicle (event)|*, *SCTID = 20902002 |Fall from bed (event)|*) within SNOMED CT corresponding to the observation table according to the types of fall [5].

The Hendrich II Fall Risk Model total score and patient acuity score were mapped to *444514002 |Hospital falls risk assessment score for the elderly (observable entity)|* and *425705009 |Determination of acuity level (procedure)|* concepts, respectively, and loaded into the observation table. Clinical observation data such as vital signs, level of consciousness (for example, Glasgow Coma Scale score), volume of output (such as urine and fluid), and pain score were also manually mapped to standard concepts within the LOINC or SNOMED CT codes corresponding to the measurement and observation tables. The LOINC and SNOMED CT codes were identified in the OHDSI standard vocabulary to arrive at OHDSI concept identifiers.

For internal and external validation, the mapping results of a nurse with abundant terminology mapping experience were reviewed and validated by a clinical terminologist. When both agreed on the mapping results, they were considered internally valid. If they disagreed, the results were discussed in group meetings attended by other researchers and domain experts (such as critical care nurses and clinical pathologists) who were not involved in the mapping process, but had experience of SNOMED CT or LOINC mapping. The measurement and observation tables were linked to *PERSON* and *VISIT_OCCURRENCE* tables based on their foreign keys (Figure 1). After completing the extraction, transformation, and load process, data quality was assessed by ACHILLES [13]. Finally, fall-related electronic health record data integrated into the existing common data model were utilized for the feasibility test.

Figure 1. Relationship between common data model tables containing fall-related electronic health record data.
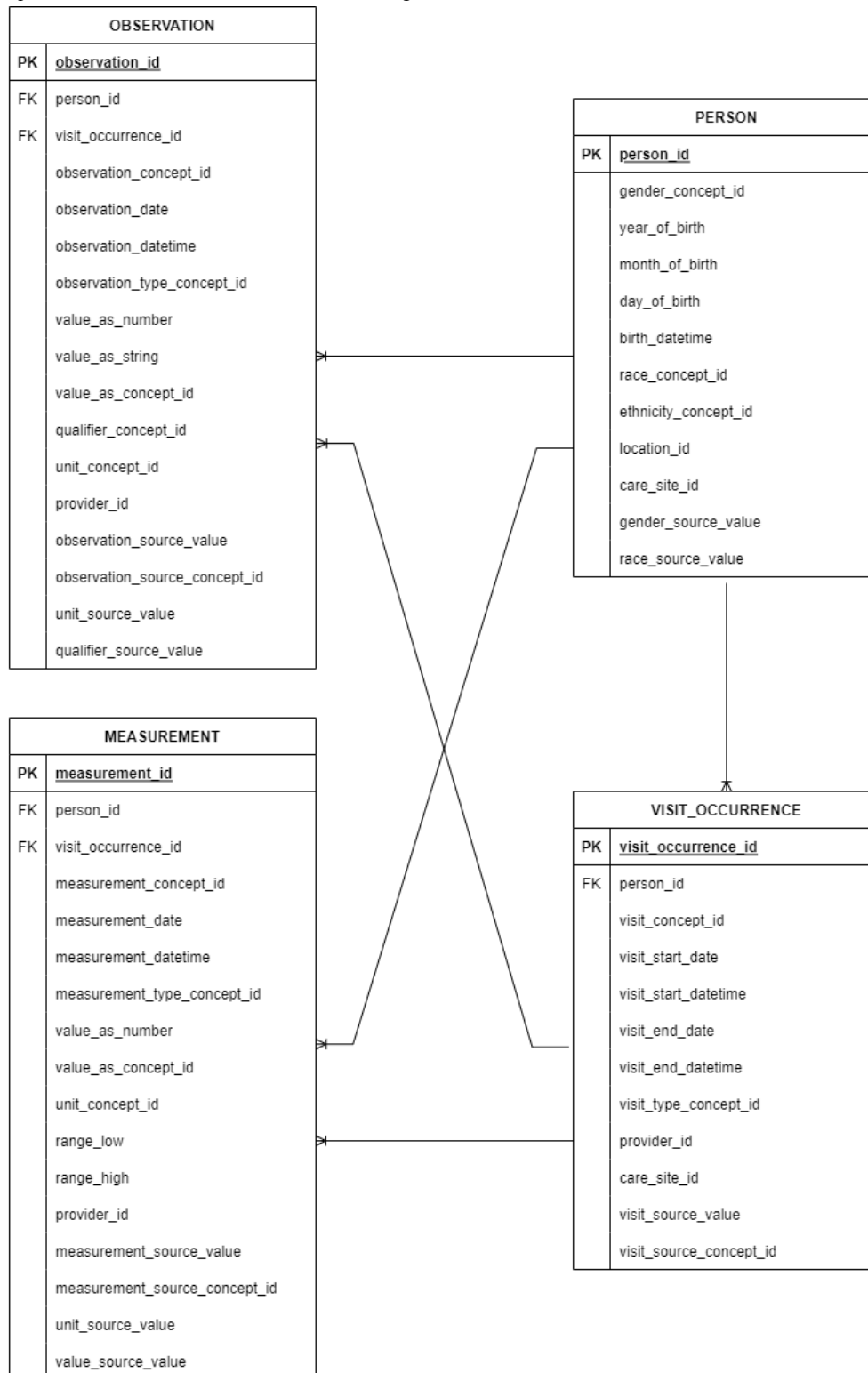
Fall Risk Prediction Using Open-Source OHDSI Analytic Tools

We utilized multiple covariates including sex, age (over 60 years of age or not), diagnoses, prescriptions, history of falling, values in laboratory tests or vital signs, interventions or surgical procedures, an atrial fibrillation stroke risk score (CHA2DS2-VASc—congestive heart failure, arterial hypertension, age over 75 years, diabetes, stroke or transient ischemic attack, vascular disease, age 65-74 years, sex), Diabetes Complications Severity Index, Charlson comorbidity score, patient acuity score, and visit count to train the prediction model.

The baseline characteristics of the study population with and without falls occurring during entire hospital stays from admission day onward were compared using 2-tailed independent sample t tests for continuous covariates and chi-square tests for categorical covariates. Furthermore, the

observation periods used to construct covariates were set as 30 and 90 days prior to the admission date (excluding the day of admission). Data recorded more than 90 days prior to admission were not included when constructing the covariates, because we considered patients' conditions at the time closest to admission to be most important in predicting the risk of falling.

The study population was randomly split into training (75%) and testing (25%) sets. We used least absolute shrinkage and selection operator–based logistic regression and random forest algorithms and selected optimal hyperparameters using 5-fold cross validation with the training set. For the evaluation of predictive performance, the area under the receiver operating characteristic curve (AUROC), sensitivity, specificity, and negative predictive value were calculated and compared to those from the Hendrich II Fall Risk Model from the same interval (within 7 days after admission or throughout the hospital stay). The total score ranges from 0 to 16, with the patient being considered at a high risk of falling when the total score is 5 or higher. Since the aim of assessing or predicting the risk of falling is to provide more effective preventive care, we gave greater weight to sensitivity and negative predictive value among the predictive indicators [14,15].

An open-source package (PatientLevelPrediction, version 4.0.5) and R software (version 4.0.3) were used to develop and evaluate the prediction model.

This study was approved by and performed in accordance with the relevant guidelines and regulations of the Seoul National University Bundang Hospital institutional review board (X-2106/689-902).

Results

Fall-Related Electronic Health Record Conversion Into OMOP Common Data Model

Of the 385,272,691 nursing statements extracted from the nursing notes, 6277 records representing fall accidents were converted to the observation table in the OMOP common data model. We converted 747,049,486 records within the clinical observation sheet into the OMOP common data model, of which 84.3% (629,535,325 records) were represented by standard vocabularies (SNOMED CT: 64,008,725/747,049,486 8.6%; LOINC: 565,526,600/747,049,486, 75.7%), and 15.7% (117,514,161/747,049,486) were not (Table 1). A total of 1,554,775 Hendrich II Fall Risk Model total scores and 5,685,011 acuity scores were converted into the observation table in OMOP common data model. Sample descriptive reports (Figures 2 and 3) from the OHDSI ACHILLES data characterization program show the prevalence of concepts per 1000 people by sex, age group, and year.

Table 1. Fall-related electronic health record data standardization.

Electronic health record data, common data model domain, and standard vocabulary	Mapped items, n	Converted records, n
Nursing statement		
Observation		
SNOMED CT	9	6277
Clinical observation sheet records		
Measurement		
LOINC	199	520,381,084
SNOMED CT	11	7,421,380
Observation		
LOINC	74	45,145,516
SNOMED CT	18	56,587,345
Fall risk score		
Observation		
SNOMED CT	1	1,554,775
Patient acuity score		
Observation		
SNOMED CT	1	5,685,011

Figure 2. Descriptive common data model report for the fall from bed concept.
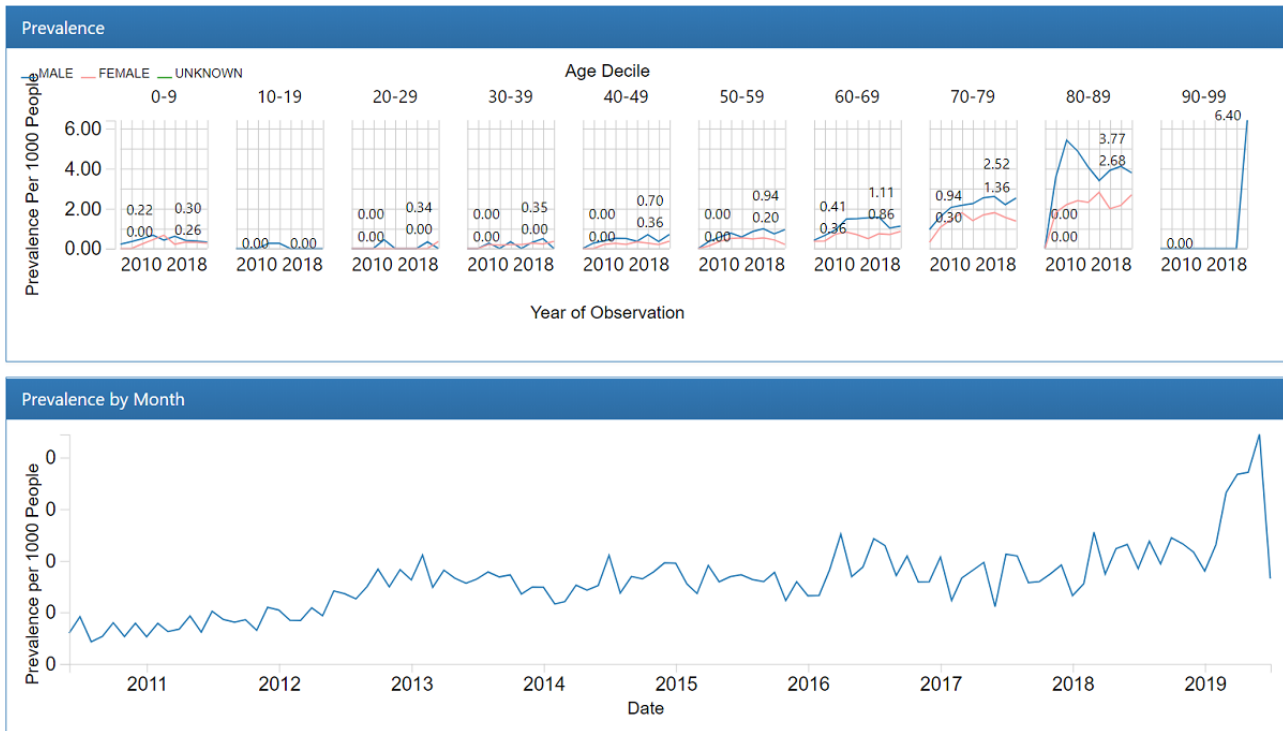
Figure 3. Descriptive common data model report (hospital falls risk assessment score for older adults).

Characteristics of the Study Population

A total of 109,289 inpatients were admitted to the neurology, neurosurgery, hematology, and oncology departments. Among them, 1465 patients fell during their hospitalization. In patients who had a fall, a larger proportion were aged 70-79 years; had previously fallen; had malignant neoplastic disease (individuals who had fallen: 753/1,465, 51.4%; individuals who had not fallen: 44,605/107,824, 41.4%), which was the most frequently observed condition medical history within 90 days before admission; and tended to take more central nervous system agents, such as antiepileptics, antidepressants, and antipsychotics

than individuals who had not fallen. At the time of admission, the mean Hendrich II Fall Risk Model total scores for individuals who had and who had not fallen were 4.3 and 2.5 points, respectively, and patient acuity score for individuals who had fallen (mean 23.2 points) was higher than that of individuals who had not fallen (mean 19.8 points). The median duration of hospital stay for individuals who had not fallen was 4 days, whereas that for individuals who had fallen was 15 days (Table 2).

Table 2. Study population.

Characteristic	Fall (n=1465)	No fall (n=107,824)	P value
Sex, n (%)			<.001
Male	845 (57.68)	56,369 (52.28)	
Female	620 (42.32)	51,455 (47.72)	
Age group (years), n (%)			
19-29	40 (2.73)	4339 (4.02)	.01
30-39	68 (4.64)	7368 (6.83)	<.001
40-49	141 (9.62)	14,457 (13.41)	<.001
50-59	231 (15.77)	25,033 (23.22)	<.001
60-69	388 (26.49)	26,857 (24.91)	.17
70-79	449 (30.65)	22,931 (21.27)	<.001
Over 80	148 (10.10)	6839 (6.34)	<.001
Previous fall, n (%)	103 (7.03)	2143 (1.99)	<.001
Condition, n (%)[a]			
Malignant neoplastic disease	753 (51.40)	44,605 (41.37)	<.001
Intracranial aneurysm	31 (2.12)	13,231 (12.27)	<.001
Neoplasm of head	234 (15.97)	9562 (8.87)	<.001
Diabetes	105 (7.17)	4328 (4.01)	<.001
Traumatic and nontraumatic brain injury	95 (6.48)	4435 (4.11)	<.001
Osteoarthritis	30 (2.05)	1042 (0.97)	<.001
Medication use, n (%)[a]			
Antiepileptics	392 (26.76)	15,639 (14.50)	<.001
Antidepressants	181 (12.35)	6969 (6.46)	<.001
Antipsychotics	188 (12.83)	6566 (6.09)	<.001
Vasoprotectives	813 (55.49)	47,284 (43.85)	<.001
Antihemorrhagics	197 (13.45)	9375 (8.69)	<.001
Procedure or operation, n (%)[a]			
Computed tomography of brain without contrast	216 (14.74)	8577 (7.95)	<.001
Magnetic resonance imaging of head and neck with contrast	86 (5.87)	5681 (5.27)	.34
Transfusion of platelet concentrate	105 (7.17)	4067 (3.77)	<.001
Measurement value, mean			
Percentage segmented neutrophils in blood	65.92	60.70	<.001
Heart rate	84.09	80.04	<.001
Glucose level (mg/dL in serum, plasma, or blood)	128.74	117.73	<.001
Visit count, mean	13.31	10.69	<.001
Hendrich II Fall Risk Score, mean	4.29	2.48	<.001
Patient acuity score, mean	23.17	19.75	<.001
Length of stay (days), median	15	4	<.001

[a]Not applicable to all patients; therefore, percentages do not add to 100%.

Predictive Performance

A total of 220 individuals who had fallen (0.81%) among 27,201 inpatients and 369 (1.36%) individuals who had fallen among 27,109 inpatients were identified to have fallen within 7 days of admission and during their entire stay, respectively, from testing set by time. The prediction feasibility test based on common data model data yielded AUROC values from 0.692

to 0.726. In general, our models showed better predictive performance than that of the Hendrich II Fall Risk Model, which used data recorded at admission or at the closest date to the admission date (Table 3). In calibration plots for the models

(Multimedia Appendix 2), confidence intervals were wide due to the low frequency of falls; nevertheless, the predicted and observed risks tended to be proportional.

Table 3. Predictive performance.

Time point and algorithm	Outcome rate (%)	AUROC (95% CI)	Sensitivity (%)	Specificity (%)	Negative predictive value (%)
Within 7 days of admission					
LASSO[a] logistic regression	0.81	0.718 (0.686-0.750)	65.91	64.24	99.57
Random forest		0.692 (0.661-0.724)	66.82	62.51	99.57
Hendrich II Fall Risk Model	0.78	0.677 (0.658-0.696)	53.81	74.52	99.51
During entire hospital stay					
LASSO logistic regression	1.36	0.726 (0.702-0.750)	68.29	63.43	99.31
Random forest		0.723 (0.698-0.747)	69.11	62.87	99.33
Hendrich II Fall Risk Model	1.35	0.673 (0.659-0.687)	52.43	74.05	99.13

[a]LASSO: least absolute shrinkage and selection operator.

Variables

Of 13,405 candidate predictors, 103 and 154 covariates were included in the logistic regression models for 7 days after admission and for the entire hospital stay, respectively. Among the top 20 covariates in the logistic regression models (Figure

4), 6 were selected for both time models. Patients' acuity scores and fall histories were the most powerful in increasing the risk of falls in the logistic regression model. In addition, age over 60 years, antiepileptic medications, C-reactive protein in serum or plasma, and erythrocyte count in body fluid were identified as common important covariates (Multimedia Appendix 3).

Figure 4. Top 20 covariates included in the logistic regression models by risk time period.
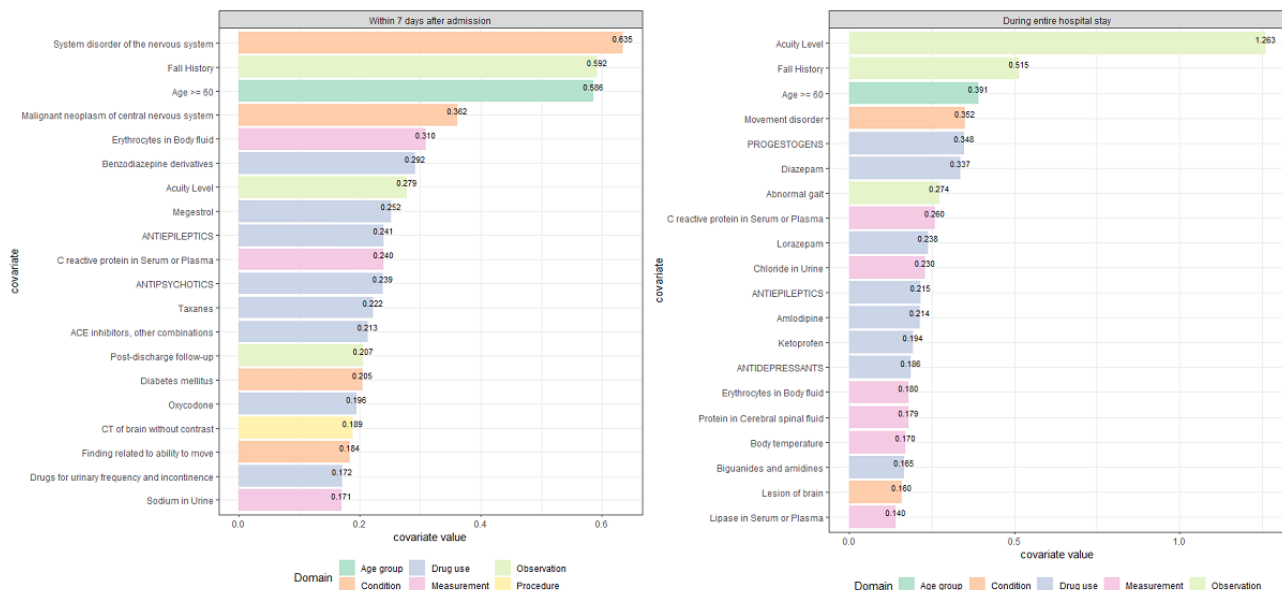
Discussion

Principal Results

To the best of our knowledge, this study is the first attempt predicting inpatients' fall risk in an acute care setting using OMOP common data model data. We converted fall-related nursing statements, fall risk, patient acuity scores, and clinical observation sheet records into OMOP common data model format using standard terminologies such as SNOMED CT and LOINC.

In the process of transforming fall-related electronic health record data to OMOP common data model format, we were able to map 50.0% (306/612) of the data items in the clinical observation sheet to the standard vocabularies, covering 84.3% (629,535,325/747,049,486) of the total clinical observation sheet records. The categories *drain* and *medication* were not mapped to the standard vocabularies. The *drain* category contained data items related to fluid output by tube type (for example, Jackson-Pratt drain, chest tube, jejunostomy, and external ventricular drain), and it is impossible to represent fluid output by tube type using predefined standard vocabularies. The *medication* category included items on volume of parenteral

fluid input by drug type (such as dopamine, herben, epinephrine, and heparin); we could not map these to more detailed precoordinated concepts than *251855004 |Parenteral fluid input (observable entity)|*.

According to some studies [16,17], typically, more than 1000 patients with the outcome being studied are needed for developing a prediction model. We found that 1465 of 109,289 inpatients had experienced accidental falls during their hospital stays—an incidence of 1.34%. These figures are lower than those of previous studies—1.66% [6] and 3.50% [18]. Boyce et al [9] showed that there was a gap in the number of falls between data sources [9]; if we had included the fall incident report as a data source, individuals who had fallen who had not been recorded in the nursing notes could have been identified.

Nursing notes are valuable sources of data for inpatient fall research. One study [19] found that nursing notes had the highest coverage for features related to inpatients' fall. For example, records of walking aid use, lower extremity strength, caregivers, fatigue, and sleep disturbance, which are important factors affecting the occurrence of falls, can only be extracted from nursing notes. Nevertheless, most nursing note records have yet to be converted into the OMOP common data model data format, and therefore, could not be used here. With the complete conversion of nursing records into the common data model data format in the near future, we expect that the performance of fall risk prediction models will be improved.

The AUROC, sensitivity, and negative predictive values were higher than those of Hendrich II Fall Risk Model assessed at the time of admission for the same patients. When we applied the best threshold (2.5 points) to Hendrich II Fall Risk Model, sensitivities improved to 62.46%-63.61% by risk time; however, this was still lower than the sensitivities (65.91%-69.11%) of the developed models. In this study, logistic regression had better predictive performance than that of random forest algorithms in terms of the AUROCs at both risk time periods, whereas the logistic regression showed similar or rather lower performance than that of random forest algorithms in terms of sensitivity and negative predictive value.

With respect to the possibility of applying prediction rules to other data, logistic regression is superior to the random forest method since the regression coefficients are known and can be applied [20]. Additionally, least absolute shrinkage and selection operator–based logistic regression models generally have greater parsimony than other machine learning models [11,21]; our study also showed this; thus, the least absolute shrinkage and selection operator–based logistic regression model is used as the reference algorithm for the OHDSI analytics pipeline for patient-level prediction.

Comparison With Prior Work

The predictive performance of our models was higher than those of another study [18], while the AUROC values were lower than those of other studies [5,6], that also used nursing records. These studies [5,6] have critical limitations, in that, they used small samples (15,480 [5] and 14,307 [6] admissions) and excluded covariates related to clinical laboratory test results and visit count. In particular, Cho et al [6] may have overfitted

by oversampling the individuals who had fallen by using the synthetic minority oversampling technique to eliminate the data imbalance between individuals who had fallen and individuals who had not fallen.

The patient acuity score was identified as the most important covariate (covariate value 1.263) for falling from the logistic regression model. Previous studies [5,22] also reported that individuals who had fallen had higher acuity scores compared to individuals who had not fallen. The patient acuity score is calculated based on clinical patient characteristics and nursing workload, which reflects patient dependency and severity. For example, the number of visits for oral medications and number of complicated intravenous drugs or transfusion are components of the patient acuity tool. Mobility or movement difficulty, which is an important factor for falling, is also reflected in the patient acuity score. Therefore, the patient acuity score could be an indicator of patients' physical vulnerability and could reflect fall risk.

Fall history was also identified as an important covariate (covariate value 0.515) for falling, which is consistent with the findings of systematic literature reviews [23,24]. American Geriatrics Society/British Geriatrics Society [25] and Australian [26] Clinical Practice Guidelines for the prevention of falls in older adults recommend that all patients be asked whether they have fallen previously because a history of falling is generally a good predictor of future falls. Additionally, having fallen within a 3-month period is one of variables of the Morse Fall Scale and Downton Fall Risk Index, which assesses a patient's likelihood of falling in an acute care setting. Patients who had experienced accidental falls feared falling [27]. It has been estimated that the prevalence of fear of falling in older adults is approximately 90% among those who had previously fallen compared to 65% among individuals who had not previously fallen [27,28]. Fear of falling in persons who have previously fallen may originate from a concern about falling, loss of balance, loss of confidence, and avoidance of activities.

The presence of movement disorders, such as abnormal gait, which has been consistently identified as a strong risk factor for falling by some systematic reviews [24,29], was also included in the top 20 covariates of fall risk prediction models using the least absolute shrinkage and selection operator–based logistic regression algorithm. Furthermore, central nervous system agents (antiepileptics and benzodiazepine derivatives, such as diazepam) were identified as predictors in this study. This result is consistent with that of a previous study [30] that suggested that psychotropic medications increased fall risk by 1.36-1.39 times in older adults. Interestingly, progestogens were identified as important predictors (covariate value 0.348) of falls that occurred for the entire hospital stay period. Although we cannot explain their causal relationship, it is well known that the levels of reproductive hormones such as estrogen and progesterone is related to the development of musculoskeletal disorders in women. As such, since some musculoskeletal disorders were identified as predictors of falls, progestogens may have indirectly influenced the occurrence of some of the falls.

Limitations

The following limitations should be considered when interpreting the results of this study. First, the generalizability of the models remains low, since the basis of the models' development was patients admitted to specific medical departments of a single hospital from 2010 to 2019. However, since the purpose of this study was to demonstrate the feasibility of the prediction model, the fall risk prediction model will be improved by adding all nursing records and applying external validation in the future. Second, the number of falls may have been underestimated, as we used only structured nursing statements to identify the occurrence of falls. Third, we only used logistic regression and random forest algorithms to develop fall prediction models, because regression-based algorithms perform better in smaller, single-center data sets [21,31] and have greater parsimony than other machine learning models [11]. In the future, we will work with multiple institutions to conduct OMOP common data model–based fall prediction research and apply other modern machine learning algorithms.

Fourth, because the conversion of electronic health record data to standard vocabularies depends on the quality of the mapping tables or the medical coder's (or terminologist's) mapping skills [32], mapping results could differ depending on mapping purpose and institutions. Therefore, since standard vocabularies change constantly, all researchers and institutions utilizing OMOP common data model data should have in-depth understanding and training on standard terminologies.

Conclusions

To the best of our knowledge, this is the first study to transform fall-related electronic health record data into OMOP common data model format and utilize the resulting data to develop fall risk prediction models for acute care settings. The performance of the developed models was superior to that of Hendrich II Fall Risk Model, which the study hospital uses to screen fall risk. Patient acuity score, history of falls, age over 60 years, movement disorder, and central nervous system agents such as psychotropic medications were identified as important covariates for fall risk prediction.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Detailed definitions of target and outcome cohorts.
[TXT File , 26 KB - medinform_v10i3e35104_app1.txt]

Multimedia Appendix 2
Calibration plots.
[PPTX File , 438 KB - medinform_v10i3e35104_app2.pptx]

Multimedia Appendix 3
Full list of covariates selected for logistic regression and random forest models by risk time period.
[XLSX File (Microsoft Excel File), 46 KB - medinform_v10i3e35104_app3.xlsx]

References

1. Dykes PC, Burns Z, Adelman J, Benneyan J, Bogaisky M, Carter E, et al. Evaluation of a patient-centered fall-prevention tool kit to reduce falls and injuries: a nonrandomized controlled trial. JAMA Netw Open 2020 Nov 02;3(11):e2025889 [FREE Full text] [doi: 10.1001/jamanetworkopen.2020.25889] [Medline: 33201236]
2. Florence CS, Bergen G, Atherly A, Burns E, Stevens J, Drake C. Medical costs of fatal and nonfatal falls in older adults. J Am Geriatr Soc 2018 Apr;66(4):693-698 [FREE Full text] [doi: 10.1111/jgs.15304] [Medline: 29512120]
3. Hendrich AL, Bender PS, Nyhuis A. Validation of the Hendrich II Fall Risk Model: a large concurrent case/control study of hospitalized patients. Appl Nurs Res 2003 Feb;16(1):9-21. [doi: 10.1053/apnr.2003.YAPNR2] [Medline: 12624858]
4. Morse JM, Morse RM, Tylko SJ. Development of a scale to identify the fall-prone patient. Can J Aging 2010 Nov 29;8(4):366-377. [doi: 10.1017/s0714980800008576]
5. Jung H, Park H, Hwang H. Improving prediction of fall risk using electronic health record data with various types and sources at multiple times. Comput Inform Nurs 2020 Mar;38(3):157-164. [doi: 10.1097/CIN.0000000000000561] [Medline: 31498252]

XSL•FO
RenderX

6.   Cho I, Boo E, Chung E, Bates DW, Dykes P. Novel approach to inpatient fall risk prediction and its cross-site validation using time-variant data. J Med Internet Res 2019 Feb 19;21(2):e11505 [FREE Full text] [doi: 10.2196/11505] [Medline: 30777849]

7.   Marier A, Olsho LEW, Rhodes W, Spector WD. Improving prediction of fall risk among nursing home residents using electronic medical records. J Am Med Inform Assoc 2016 Mar;23(2):276-282. [doi: 10.1093/jamia/ocv061] [Medline: 26104743]

8.   Reps JM, Schuemie MJ, Suchard MA, Ryan PB, Rijnbeek PR. Design and implementation of a standardized framework to generate and evaluate patient-level prediction models using observational healthcare data. J Am Med Inform Assoc 2018 Aug 01;25(8):969-975 [FREE Full text] [doi: 10.1093/jamia/ocy032] [Medline: 29718407]

9.   Boyce RD, Handler SM, Karp JF, Perera S, Reynolds CF. Preparing nursing home data from multiple sites for clinical research - a case study using observational health data sciences and informatics. EGEMS (Wash DC) 2016;4(1):1252 [FREE Full text] [doi: 10.13063/2327-9214.1252] [Medline: 27891528]

10. Jin S, Kostka K, Posada JD, Kim Y, Seo SI, Lee DY, et al. Prediction of major depressive disorder following beta-blocker therapy in patients with cardiovascular diseases. J Pers Med 2020 Dec 18;10(4):288 [FREE Full text] [doi: 10.3390/jpm10040288] [Medline: 33352870]

11. Wang Q, Reps JM, Kostka KF, Ryan PB, Zou Y, Voss EA, et al. Development and validation of a prognostic model predicting symptomatic hemorrhagic transformation in acute ischemic stroke at scale in the OHDSI network. PLoS One 2020;15(1):e0226718 [FREE Full text] [doi: 10.1371/journal.pone.0226718] [Medline: 31910437]

12. Kim J, Kim S, Ryu B, Song W, Lee H, Yoo S. Transforming electronic health record polysomnographic data into the Observational Medical Outcome Partnership's common data model: a pilot feasibility study. Sci Rep 2021 Mar 29;11(1):7013 [FREE Full text] [doi: 10.1038/s41598-021-86564-w] [Medline: 33782494]

13. The common data model. The Book of OHDSI. 2021 Jan 11. URL: https://ohdsi.github.io/TheBookOfOhdsi/ [accessed 2022-03-02]

14. Jung H, Park H. Testing the predictive validity of the Hendrich II fall risk model. West J Nurs Res 2018 Dec;40(12):1785-1799. [doi: 10.1177/0193945918766554] [Medline: 29577823]

15. Lee Y, Jeong I, Jeon S. A comparative study on the predictive validity among pressure ulcer risk assessment scales. Taehan Kanho Hakhoe Chi 2003 Apr;33(2):162-169. [doi: 10.4040/jkan.2003.33.2.162] [Medline: 15314444]

16. John L, Kors J, Reps J, Ryan P, Rijnbeek P. How little data do we need for patient-level prediction? Arxiv. Preprint posted online Aug 14, 2020. [FREE Full text] [doi: 10.48550/arXiv.2008.07361]

17. Khalid S, Yang C, Blacketer C, Duarte-Salles T, Fernández-Bertolín S, Kim C, et al. A standardized analytics pipeline for reliable and rapid development and validation of prediction models using observational health data. Comput Methods Programs Biomed 2021 Nov;211:106394 [FREE Full text] [doi: 10.1016/j.cmpb.2021.106394] [Medline: 34560604]

18. Yokota S, Ohe K. Construction and evaluation of FiND, a fall risk prediction model of inpatients from nursing data. Jpn J Nurs Sci 2016 Apr;13(2):247-255. [doi: 10.1111/jjns.12103] [Medline: 27040735]

19. Jung H, Park H. Use of EHR Data to Identify Factors Affecting the Time to Fall. Stud Health Technol Inform 2017;245:1043-1047. [Medline: 29295260]

20. Couronné R, Probst P, Boulesteix A. Random forest versus logistic regression: a large-scale benchmark experiment. BMC Bioinformatics 2018 Jul 17;19(1):270 [FREE Full text] [doi: 10.1186/s12859-018-2264-5] [Medline: 30016950]

21. Christodoulou E, Ma J, Collins GS, Steyerberg EW, Verbakel JY, Van Calster B. A systematic review shows no performance benefit of machine learning over logistic regression for clinical prediction models. J Clin Epidemiol 2019 Jun;110:12-22. [doi: 10.1016/j.jclinepi.2019.02.004] [Medline: 30763612]

22. Hong H, Kim N, Jin Y, Piao J, Lee S. Trigger factors and outcomes of falls among korean hospitalized patients: analysis of electronic medical records. Clin Nurs Res 2015 Feb;24(1):51-72. [doi: 10.1177/1054773814524225] [Medline: 24615824]

23. Oliver D, Daly F, Martin FC, McMurdo MET. Risk factors and risk assessment tools for falls in hospital in-patients: a systematic review. Age Ageing 2004 Mar;33(2):122-130. [doi: 10.1093/ageing/afh017] [Medline: 14960426]

24. Ambrose AF, Paul G, Hausdorff JM. Risk factors for falls among older adults: a review of the literature. Maturitas 2013 May;75(1):51-61. [doi: 10.1016/j.maturitas.2013.02.009] [Medline: 23523272]

25. Panel on Prevention of Falls in Older Persons, American Geriatrics Society/British Geriatrics Society. Summary of the updated American Geriatrics Society/British Geriatrics Society clinical practice guideline for prevention of falls in older persons. J Am Geriatr Soc 2011 Jan;59(1):148-157. [doi: 10.1111/j.1532-5415.2010.03234.x] [Medline: 21226685]

26. Australian Commission on Safety and Quality in Healthcare. Preventing falls and harm from falls in older people: best practice guidelines for Australian hospitals. Australian Clinical Practice Guidelines. URL: https://www.safetyandquality.gov.au/sites/default/files/migrated/Guidelines-HOSP.pdf [accessed 2022-03-02]

27. Jørstad EC, Hauer K, Becker C, Lamb SE, ProFaNE Group. Measuring the psychological outcomes of falling: a systematic review. J Am Geriatr Soc 2005 Mar;53(3):501-510. [doi: 10.1111/j.1532-5415.2005.53172.x] [Medline: 15743297]

28. Gazibara T, Kurtagic I, Kisic-Tepavcevic D, Nurkovic S, Kovacevic N, Gazibara T, et al. Falls, risk factors and fear of falling among persons older than 65 years of age. Psychogeriatrics 2017 Jul;17(4):215-223. [doi: 10.1111/psyg.12217] [Medline: 28130862]

29. Moreland JD, Richardson JA, Goldsmith CH, Clase CM. Muscle weakness and falls in older adults: a systematic review and meta-analysis. J Am Geriatr Soc 2004 Jul;52(7):1121-1129. [doi: 10.1111/j.1532-5415.2004.52310.x] [Medline: 15209650]

30. Leipzig RM, Cumming RG, Tinetti ME. Drugs and falls in older people: a systematic review and meta-analysis: I. Psychotropic drugs. J Am Geriatr Soc 1999 Jan;47(1):30-39. [doi: 10.1111/j.1532-5415.1999.tb01898.x] [Medline: 9920227]

31. Sanchez-Pinto LN, Venable LR, Fahrenbach J, Churpek MM. Comparison of variable selection methods for clinical predictive modeling. Int J Med Inform 2018 Aug;116:10-17 [FREE Full text] [doi: 10.1016/j.ijmedinf.2018.05.006] [Medline: 29887230]

32. Reich C, Ryan PB, Stang PE, Rocca M. Evaluation of alternative standardized terminologies for medical conditions within a network of observational healthcare databases. J Biomed Inform 2012 Aug;45(4):689-696 [FREE Full text] [doi: 10.1016/j.jbi.2012.05.002] [Medline: 22683994]

Abbreviations

ACHILLES: Automated Characterization of Health Information at Large-scale Longitudinal Evidence Systems
AUROC: area under the receiver operating characteristics curve
LOINC: Logical Observation Identifiers Names and Codes
OHDSI: Observational Health Data Science and Informatics
OMOP: Observational Medical Outcome Partnership
SNOMED CT: Systematized Nomenclature of Medicine–Clinical Terms

<u>Original Paper</u>

Foundations for Meaningful Consent in Canada's Digital Health Ecosystem: Retrospective Study

Nelson Shen[1,2], MHA, PhD; Iman Kassam[1], BSc, MHI; Haoyu Zhao[1], MPH; Sheng Chen[1], PhD; Wei Wang[1,3], PhD; Sarah Wickham[4], BSc; Gillian Strudwick[1,2], RN, PhD; Abigail Carter-Langford[4], LLM

[1]Centre for Complex Interventions, Centre for Addiction and Mental Health, Toronto, ON, Canada

[2]Institute of Health Policy, Management and Evaluation, University of Toronto, Toronto, ON, Canada

[3]College of Public Health, University of South Florida, Tampa, FL, United States

[4]Canada Health Infoway, Toronto, ON, Canada

Corresponding Author:
Nelson Shen, MHA, PhD
Centre for Complex Interventions
Centre for Addiction and Mental Health
60 White Squirrel Way
Toronto, ON, M6J 1H4
Canada
Phone: 1 416 535 8501
Email: nelson.shen@camh.ca

Abstract

Background: Canadians are increasingly gaining web-based access to digital health services, and they expect to access their data from these services through a central patient access channel. Implementing data sharing between these services will require patient trust that is fostered through meaningful consent and consent management. Understanding user consent requirements and information needs is necessary for developing a trustworthy and transparent consent management system.

Objective: The objective of this study is to explore consent management preferences and information needs to support meaningful consent.

Methods: A secondary analysis of a national survey was conducted using a retrospective descriptive study design. The 2019 cross-sectional survey used a series of vignettes and consent scenarios to explore Canadians' privacy perspectives and preferences regarding consent management. Nonparametric tests and logistic regression analyses were conducted to identify the differences and associations between various factors.

Results: Of the 1017 total responses, 716 (70.4%) participants self-identified as potential users. Of the potential users, almost all (672/716, 93.8%) felt that the ability to control their data was important, whereas some (385/716, 53.8%) believed that an *all or none* control at the data source level was adequate. Most potential users preferred new data sources to be accessible by health care providers (546/716, 76.3%) and delegated parties (389/716, 54.3%) by default. Prior digital health use was associated with greater odds of granting default access when compared with no prior use, with the greatest odds of granting default access to digital health service providers (odds ratio 2.17, 95% CI 1.36-3.46). From a list of 9 information elements found in consent forms, potential users selected an average of 5.64 (SD 2.68) and 5.54 (SD 2.85) items to feel informed in consenting to data access by care partners and commercial digital health service providers, respectively. There was no significant difference in the number of items selected between the 2 scenarios ($P > .05$); however, there were significant differences ($P < .05$) in information types that were selected between the scenarios.

Conclusions: A majority of survey participants reported that they would register and use a patient access channel and believed that the ability to control data access was important, especially as it pertains to access by those outside their care. These findings suggest that a broad *all or none* approach based on data source may be accepted; however, approximately one-fifth of potential users were unable to decide. Although vignettes were used to introduce the questions, this study showed that more context is required for potential users to make informed consent decisions. Understanding their information needs will be critical, as these needs vary with the use case, highlighting the importance of prioritizing and tailoring information to enable meaningful consent.

XSL·FO
RenderX

Introduction

Background

Canadians are becoming increasingly aware of digital health tools and services to support their health and wellness and are beginning to demand that they have greater access to their data that are held within these tools and services. Those who accessed their health records reported that they were more knowledgeable, informed, and confident about the care they received [1,2]. Although there are benefits to having a wide variety of digital health tools and services available, the rapid growth of the digital health ecosystem has resulted in silos of patient data. The prospect of universally connecting digital health tools, such as patient portals, is a challenge, given the large number of data exchange protocols required to share information between all points in a patient's journey [3]. Historically, patient portals have been implemented at the organizational level and tethered to their organizational electronic health record (EHR) system. As these portals seldom exchange information between organizations, patients may end up with multiple portals of siloed data based on the various points where they seek care [4]. As a result, many patients have fragmented, limited, or no electronic access to their personal health information (PHI), giving patients an incomplete picture of their overall health to support their health care decisions. Furthermore, the multiplicity of tools and services may provide an additional burden to patients as they will need to manage the different log-ins and privacy preferences for each one.

There are growing patient demands and expectations for web-based access to their consolidated clinical and self-generated data through a single access point, recognizing that it will *make their lives better* [5]. A patient access channel serves as a trusted access point, granting patients authenticated access to their PHI and digital services data within a single platform. This allows patients to manage the collection, use, and disclosure of their PHI. Patients have the right to control how their information is collected and used, which is the definition of information privacy [6]. Canadian legislative frameworks provide protection and, generally, enable individuals to limit the use and disclosure of their records to certain individuals for specific purposes [7]. Implementing a consent management system would empower users to exercise their data-sharing preferences [8,9].

Privacy Notices and Consent

Canadian legislation also requires consent for the collection, use, and disclosure of personal information and PHI; however, consent is seldom transparent or informed, leaving patients unaware of how their data are used and with minimal control over their data [10]. Given the largely unregulated commercial digital health ecosystem, digital health services are founded in a business model where user data are often sold for marketing or other purposes that the user may not be able to understand or foresee [11,12]. In these contexts, consent is illusory and a form of *coercion* as it does not reflect informed

choice—individuals are left with the ultimatum to use or not with minimal understanding of what they are consenting to [13]. On average, privacy notices are 3964 words in length and take 18 minutes to read [14]; moreover, they are written at a postuniversity level [15]. There is an ethical imperative to improve the transparency of data use and user control of data to avoid any future exploitation by entities collecting the data [16,17].

The patient access channel offers the potential to implement consent standards that enable transparent and meaningful consent. The Office of the Privacy Commissioner of Canada Meaningful Consent Guidelines include actionable recommendations for organizations to strengthen their digital consent practices [18] by:

1. Emphasizing key elements
2. Allowing individuals to control the level of detail they get and when
3. Providing individuals with clear options to say *yes* or *no*
4. Being innovative and creative
5. Considering the consumer's perspective
6. Making consent a dynamic and ongoing process
7. Being accountable and standing ready to demonstrate compliance

Although Meaningful Consent Guidelines provide a set of heuristics to improve consent processes, they are not specific to the digital health context [18]; moreover, they are only recommendations and do not require vendor compliance. The success of digital health requires trust and transparency in data use [19-21]. With privacy and trust as 2 intertwined antecedents to technology use and data-sharing behaviors [22], where their absence negatively affects use and behaviors, it is critical to understand the patient's expectations of privacy to foster trust, acceptance, and use.

Objective

A 2-stage stakeholder engagement project was conducted by Canada Health Infoway to explore the user consent requirements of a patient access channel and the privacy considerations of its implementation. It consisted of a pan-Canadian survey and regional stakeholder workshops across Canada [23]. The study reported here is a retrospective analysis of the survey data. The objective of this retrospective study is to provide a more granular understanding of user preferences for consent management.

Methods

Study Design

This retrospective study uses data from a cross-sectional national web-based survey conducted between October 2 and October 15, 2019, by Canada Health Infoway. This study explored how consent management preferences and information needs differ across various patient characteristics. Specifically, this study asked the following research questions (RQs):

XSL•FO

RenderX

- RQ1: What are the data control and consent management preferences of potential patient access channel users?
- RQ2: How do information needs differ among individuals when making an informed decision to share their health data with different individuals or entities?

Data Collection

The survey comprised a series of hypothetical vignettes and consent scenarios to solicit participants' perspectives on the consent management service and its functionalities through a mix of closed-and open-ended questions (see Multimedia Appendix 1 for the detailed vignettes and consent scenarios). There were four sections to the survey: (1) participant characteristics, (2) intention to register for the consent management service, (3) consent management use case scenarios, and (4) demographics.

The survey was administered electronically by a Canadian marketing research firm (Leger Marketing) to its pan-Canadian web panel. Using their pan-Canadian web panel, a 20-minute web-based survey was administered to the general Canadian population, reaching across the 10 provinces. The sampling strategy focused on potential digital health service users (ie, those with frequent interactions with the health care system) and used a proportional quota sampling strategy to recruit equal proportions of adults and older adults, with quotas set at 50% for adults and 80% for older adults with at least one chronic condition. The surveys were made available in English and French. Participants were eligible to participate in the survey if they were Canadian citizens, aged ≥18 years, currently live in Canada, and were within the provincial quotas for adults and seniors with chronic conditions. The survey had a view rate of 16.67% (1666/9997) and a completion rate of 61.04% (1017/1666).

Measures

Overview

This study's analytic frame comprises potential users of the patient access channel. Potential users were defined as participants who indicated that they would register to use a hypothetical patient access channel in the first set of vignettes. The vignette presented information about Canada Health Infoway and the functionality of the patient access channel (or *gateway*). Participants were then asked how likely they would register for the gateway using a 4-point Likert scale (ranging from not at all likely to very likely). Participants were also provided with an *I don't know* option throughout the survey. The second vignette introduced a *trust framework* as the *rules of operation and participation, such as policies and agreements around data sharing and how users can control their health information*. It also presents information on consent management, single sign-on, and privacy safeguards. Participants were then asked how likely it was that they would register for the gateway based on their understanding of the trust framework and the availability of safeguards. Participants who answered *somewhat* or *very* likely were categorized as potential users.

User characteristics (ie, demographics and user experiences) were used as covariates in the analysis. The variables that exhibited a low frequency of response for some scale points were collapsed into categories to improve the statistical power of the analysis [24]. Sociodemographic data included sex (male and female), age (18-44 years, 45-64 years, and ≥65 years), income (>CAD $80,000 [US $62,380] and <CAD $80,000 [US $62,380]), and region (Atlantic, Central, Prairies, and West Coast). User experiences comprised health care use (high users or low users), patient engagement (engaged or not engaged), digital health user (user or nonuser), perceived quality of care (good or poor), past web-based experiences (good or poor), health care privacy experiences (good or poor), past privacy breaches (no past breach, breach resolved, or breach not resolved), perceived confidentiality of PHI (private or not private), perceived sensitivity of PHI (high sensitivity or low sensitivity), and perceived sensitivity of digital health data (high sensitivity or low sensitivity). A median cutoff was used to establish the threshold for perceived sensitivity variables as the categories had no theoretical grounding or frame of reference. Further details about the outcome variables and covariates can be found in Multimedia Appendix 2. The full survey can be found in Multimedia Appendix 3.

There are four variables of interest in this study: (1) the importance of consent management, (2) adequacy of broad consent, (3) entities with default access to user data, and (4) user information needs to make an informed decision about data sharing.

Consent Management Preferences

Participants were presented with a vignette about privacy controls and the gateway function of enabling consent directives to block or restrict access to their PHI. Participants were then asked to rate the *importance of having the ability to change privacy preferences for sharing PHI* on a 4-point ordinal scale (*not at all important* to *very important*).

The next vignette presented a scenario regarding broad consent, where a data recipient would receive either *all or none* of a particular data source (eg, medical history, laboratory records, clinical and diagnostics, and e-service data). Participants were asked to assess whether the broad access control reflected their needs or did not reflect their needs or if they did not know.

For default access, participants were presented with a scenario where they enrolled in a new digital health service and were asked to select the entities to whom they would grant default access to new sources of data. Given that they still had the ability to apply consent directives, they were asked to select the following entities to whom they would grant default access to the new source of information (ie, select all that apply): health care providers, authorized members (ie, family and friends), digital health services and tools, or none of the above (ie, grant access individually or to each group).

Information Needs

To assess user information needs for informed consent, participants were first presented a vignette on consent management, which outlined the types of PHI they may access in the gateway and introduced an access control function that allows patients to authorize access to their PHI to health care providers, family and friends, and digital service vendors.

Participants were asked to select the types of information they required to make an informed decision on whether to share their data in two scenarios: sharing with friends and family (scenario 1) and sharing with digital health providers for the digital health service (scenario 2).

Participants were provided with a list of information types that are found on consent forms and privacy notices and were asked to select all that applied. The nine information types were as follows: what types of information that the digital service can access, what the digital service can do with their data, potential risks and benefits of granting access, how to ask more questions about information sharing or privacy, how to file complaints about how information is shared, functions that allow them to monitor activity, types of data access controls available, and how to revoke access.

Data Analysis

First, the frequencies and percentages of the characteristics and demographics of all potential users were reported. For RQ1, frequencies for the importance of access control, adequacy of *all or none* access control based on data source, and default access to PHI were shown. Logistic regression was applied to evaluate the factors associated with the adequacy of access control, whether knowing it met their needs regarding adequacy, and granting default access. In the model-building procedure, a small subset of participants was excluded from the total sample because of the limited number of observations within each cell. The number of participants and the corresponding percentages were reported for the frequency analysis. Adjusted odds ratios (ORs) and 95% CIs were reported for logistic regression results.

For RQ2, the Friedman test was used to assess the difference in the number of items selected between the 2 scenarios in terms of sharing their information. The McNemar test was also performed to check if the frequency of each item differed between the 2 scenarios. All statistical analyses were conducted using SAS software (SAS Enterprise Guide 7.1; SAS Institute Inc).

Ethics Approval

This study was approved by the Research and Ethics Board at the Centre for Addiction and Mental Health (REB#114/2020) in Toronto, Canada.

Results

Overall Results

Of the 1017 responses, 716 (70.4%) *potential users* of the patient access channel were identified. The potential user characteristics can be found in Table 1. Over three-quarters were low service users (559/716, 78.1%), noncaregivers (621/716, 86.7%), engaged patients (612/716, 85.5%), and satisfied with their quality of care (609/716, 85.1%). Over half had used digital health tools previously (471/716, 65.8%) and rated their PHI (364/716, 50.8%) and digital health data as sensitive (423/716, 59.1%). Most potential users reported having positive privacy experiences on the web (535/716, 74.7%), positive health care privacy experiences (643/716, 89.8%), and trust in the confidentiality of their records in the health care system (644/716, 89.9%). The final sample size of potential users for the logistic regression model was 712.

Table 1. Characteristics of potential users (N=716).

Characteristic	Values, n (%)
Sex	
Female	343 (47.9)
Male	369 (51.5)
Transgender[a]	2 (0.3)
Other[a]	1 (0.1)
PNA[a,b]	1 (0.1)
Age (years)	
18-44	204 (28.5)
45-64	155 (21.7)
≥65	357 (49.9)
Region	
Atlantic	47 (6.6)
Central	418 (58.4)
Prairie	145 (20.3)
West Coast	106 (14.8)
Income (CAD$; US $)	
<$80,000 ($62,380)	401 (56)
>$80,000 ($62,380)	258 (36)
PNA	57 (8)
Health care use	
High (>20)	146 (20.4)
Low (≤20)	559 (78.1)
IDK[c]	11 (1.5)
Caregiver	
No	621 (86.7)
Yes	95 (13.3)
Quality of care	
Good	609 (85.1)
Poor	92 (12.9)
IDK	15 (2.1)
Prior digital health use	
Yes	471 (65.8)
No	245 (34.2)
Engaged patient	
Yes	612 (85.5)
No	104 (14.5)
Sensitivity of PHI[d]	
High (≥10)	364 (50.8)
Low (<10)	352 (49.2)
Sensitivity of digital health data	
High (≥11)	383 (53.5)

Characteristic	Values, n (%)
Low (<11)	333 (46.5)
Web-based privacy experience	
Good	535 (74.7)
Poor	134 (18.7)
IDK	47 (6.6)
Privacy breach	
Yes, resolved	70 (9.8)
Yes, not resolved or IDK	29 (4.1)
No breach	617 (86.2)
Health care privacy experiences	
Good	643 (89.8)
Poor	51 (7.1)
IDK	22 (3.1)
Confidentiality of records	
Private	644 (89.9)
Not private	37 (5.2)
IDK	35 (4.9)

[a]Indicates subpopulations that were excluded from the logistic regression model.

[b]PNA: prefer not to answer.

[c]IDK: I do not know.

[d]PHI: personal health information.

RQ1: What Are the Data Control and Consent Management Preferences of Potential Patient Access Channel Users?

Importance of Access Control

Overall, 93.8% (672/716) of the potential users believed it was important (126/716, 18%) or very important (543/716, 75.8%) to have the ability to control their privacy preferences. Further subanalyses were not conducted as the distribution of responses would not allow for the detection of differences between the options.

Adequacy of All or None Access Control Based on Data Source

Approximately 53.8% (385/716) of the potential users felt that an *all or none* approach based on the data source to control data access was adequate for their needs, whereas 29.2% (209/716) did not, and 17.0% (122/716) did not know.

Geographic location and income were the only factors that were significantly associated with *all or none* being adequate for the participant's needs. Potential users from the Prairies were 50% less likely than those from Central Canada to feel that it was adequate (OR 0.50, 95% CI 0.32-0.78). Potential users earning >CAD $80,000 (US $62,380) or potential users that did not disclose their income were 42% and 68% less likely to find *all or none* adequate than low-income earners (<CAD $80,000 [US $62,380]; OR 0.58, 95% CI 0.40-0.86; OR 0.32, 95% CI 0.16-0.66). Potential users with high income were 129% more likely to know that an *all or none* approach would meet their needs than those with low income (OR 2.29, 95% CI 1.36-3.83). Those who used digital health tools previously were associated with a 109% increased likelihood to know that an *all or none* approach would meet their needs than those who did not (OR 2.09, 95% CI 1.34-3.25). The results of the logistic regression analysis can be found in Table 2.

Table 2. Comparison of the adequacy of all or none based on participant characteristics.

Characteristics	Adequate for needs, odds ratio (95% CI)	Know versus not know, odds ratio (95% CI)
Sex		
Male	Reference	Reference
Female	0.77 (0.53-1.13)	0.72 (0.47-1.11)
Age (years)		
18-45	Reference	Reference
46-64	0.80 (0.48-1.34)	1.09 (0.55-2.15)
≥65	0.74 (0.47-1.17)	0.68 (0.39-1.17)
Health care use		
Low	Reference	Reference
High	1.33 (0.83-2.13)	0.67 (0.40-1.12)
IDK[a]	0.99 (0.17-5.58)	0.38 (0.09-1.62)
Region		
Central	Reference	Reference
Atlantic	0.54 (0.26-1.10)	1.45 (0.56-3.77)
Prairie	0.50 (0.32-0.78)[b]	0.88 (0.52-1.50)
West Coast	0.74 (0.43-1.25)	0.59 (0.34-1.05)
Caregiver		
No	Reference	Reference
Yes	1.60 (0.93-2.74)	2.00 (0.93-4.30)
Income (CAD $; US $)		
<$80,000 ($62,380)	Reference	Reference
>$80,000 ($62,380)	0.58 (0.40-0.86)[b]	2.29 (1.36-3.83)[b]
PNA[c]	0.32 (0.16-0.66)[b]	0.79 (0.40-1.55)
Engaged patient		
No	Reference	Reference
Yes	0.92 (0.53-1.61)	0.86 (0.44-1.69)
Quality of care		
Poor	Reference	Reference
Good	0.92 (0.49-1.76)	1.45 (0.73-2.87)
IDK	2.18 (0.38-12.56)	1.26 (0.27-5.99)
Prior digital health use		
No	Reference	Reference
Yes	0.75 (0.50-1.13)	2.09 (1.34-3.25)[b]
Sensitivity of PHI[d]		
Low	Reference	Reference
High	0.68 (0.44-1.05)	1.07 (0.65-1.77)
Sensitivity of health data		
Low	Reference	Reference
High	1.25 (0.81-1.92)	0.70 (0.42-1.17)
Web-based privacy experiences		
Poor	Reference	Reference

Characteristics	Adequate for needs, odds ratio (95% CI)	Know versus not know, odds ratio (95% CI)
Good	1.27 (0.77-2.08)	1.18 (0.66-2.11)
IDK	1.28 (0.50-3.27)	0.52 (0.21-1.25)
Past privacy breach		
Not resolved or IDK	Reference	Reference
Resolved	1.22 (0.44-3.40)	1.66 (0.43-6.42)
No breaches	1.53 (0.63-3.74)	0.97 (0.32-2.97)
Health care privacy experiences		
Poor	Reference	Reference
Good	1.73 (0.81-3.71)	0.54 (0.20-1.48)
IDK	0.64 (0.13-3.12)	0.27 (0.06-1.19)
Confidentiality of PHI		
Not private	Reference	Reference
Private	0.95 (0.38-2.35)	1.09 (0.40-2.92)
IDK	1.90 (0.50-7.27)	0.87 (0.25-2.99)

[a]IDK: I do not know.

[b]Signifies a significant association when compared with the reference group.

[c]PNA: prefer not to answer.

[d]PHI: personal health information.

Default Access to PHI

Most potential users would grant default access to new data that become available to their health care providers (546/716, 76.3%) or authorized members, such as family, friends, and other care partners (389/716, 54.3%). Approximately one-fifth would grant default access to their digital health service provider for use with digital health services (138/716, 19.3%). Finally, 14.8% (106/716) of the potential users would not grant default access to anyone. Factors associated with granting default access were prior digital health use, health care privacy experiences, caregiver status, sex, and perceived sensitivity of PHI (Table 3).

Prior use of digital health tools was associated with a greater likelihood of granting default access to the 3 entities as there was a 66% greater likelihood of granting default access to health care providers (OR 1.66, 95% CI 1.14-2.44), 101% greater likelihood of granting default access to authorized members (OR 2.01, 95% CI 1.43-2.81), and 117% greater likelihood of granting default access to digital health service providers (OR 2.17, 95% CI 1.36-3.46). Those with prior digital health tool use were 53% less likely to not want to grant default access to anyone (OR 0.47, 95% CI 0.29-0.74). Those with positive health care privacy experiences were 156% more likely to grant default access to health care providers (OR 2.56, 95% CI 1.24-5.29) and 70% less likely to not grant default access than those with poor experiences (OR 0.30, 95% CI 0.20-0.70).

Service providers were 142% more likely to gain default access from caregivers (OR 2.42, 95% CI 1.45-4.04) but 39% less likely to gain default access from females (OR 0.61, 95% CI 0.40-0.92). Authorized users were 34% less likely to gain default access (OR 0.66, 95% CI 0.46-0.96) from potential users who had high perceived PHI sensitivity in comparison with those with low perceived PHI sensitivity. Those with high PHI sensitivity were also 126% more likely to not grant default access to anyone (OR 2.26, 95% CI 1.30-3.93) than those with low perceived PHI sensitivity.

Table 3. Comparison of default access based on participant characteristics.

Characteristics	Odds ratio (95% CI)			
	HCP[a]	Authorized members	DHSP[b]	No one
Sex				
Male	Reference	Reference	Reference	Reference
Female	1.12 (0.77-1.63)	0.94 (0.68-1.29)	0.61 (0.40-0.92)[c]	0.90 (0.57-1.41)
Age (years)				
18-44	Reference	Reference	Reference	Reference
45-64	0.85 (0.51-1.43)	1.31 (0.84-2.06)	0.94 (0.54-1.64)	1.25 (0.65-2.39)
≥65	0.87 (0.55-1.37)	1.35 (0.92-1.99)	0.90 (0.55-1.47)	1.39 (0.78-2.47)
Health care use				
Low (≤20)	Reference	Reference	Reference	Reference
High (>20)	1.33 (0.82-2.17)	0.79 (0.54-1.18)	0.99 (0.60-1.63)	1.04 (0.58-1.85)
IDK[d]	0.43 (0.12-1.55)	0.81 (0.22-2.96)	0.43 (0.05-3.73)	2.16 (0.48-9.73)
Region				
Central	Reference	Reference	Reference	Reference
Atlantic	0.85 (0.41-1.73)	1.13 (0.60-2.15)	1.38 (0.62-3.07)	0.98 (0.41-2.35)
Prairie	1.31 (0.80-2.15)	0.72 (0.48-1.06)	0.72 (0.42-1.23)	0.61 (0.32-1.16)
West Coast	0.72 (0.44-1.18)	0.82 (0.52-1.28)	0.93 (0.52-1.65)	1.68 (0.95-2.97)
Caregiver				
No	Reference	Reference	Reference	Reference
Yes	0.75 (0.45-1.25)	1.42 (0.89-2.27)	2.42 (1.45-4.04)[c]	0.56 (0.26-1.19)
Income (CAD $; US $)				
<$80,000 ($62,380)	Reference	Reference	Reference	Reference
≥$80,000 ($62,380)	0.89 (0.60-1.31)	1.23 (0.88-1.72)	1.26 (0.83-1.91)	1.20 (0.74-1.95)
PNA[e]	0.73 (0.38-1.40)	0.75 (0.42-1.34)	0.65 (0.26-1.63)	1.77 (0.86-3.68)
Engaged patient				
No	Reference	Reference	Reference	Reference
Yes	0.95 (0.54-1.66)	1.14 (0.70-1.84)	0.63 (0.35-1.14)	1.29 (0.63-2.65)
Quality of care				
Poor	Reference	Reference	Reference	Reference
Good	1.07 (0.58-1.98)	0.87 (0.51-1.49)	0.75 (0.39-1.46)	0.79 (0.38-1.65)
IDK	1.27 (0.29-5.59)	0.33 (0.09-1.16)	1.54 (0.36-6.56)	0.75 (0.12-4.58)
Prior digital health use				
No	Reference	Reference	Reference	Reference
Yes	1.66 (1.14-2.44)[c]	2.01 (1.43-2.81)[c]	2.17 (1.36-3.46)[c]	0.47 (0.29-0.74)[c]
Sensitivity of PHI[f]				
Low (<10)	Reference	Reference	Reference	Reference
High (≥10)	0.93 (0.60-1.43)	0.66 (0.46-0.96)[c]	1.04 (0.65-1.69)	2.26 (1.30-3.93)[c]
Sensitivity of health data				
Low (<11)	Reference	Reference	Reference	Reference
High (≥11)	0.91 (0.59-1.40)	1.16 (0.80-1.68)	1.13 (0.70-1.83)	0.87 (0.51-1.49)

XSL•FO
RenderX

Characteristics	Odds ratio (95% CI)			
	HCP[a]	Authorized members	DHSP[b]	No one
Web-based privacy experiences				
Poor	Reference	Reference	Reference	Reference
Good	0.79 (0.47-1.33)	1.03 (0.66-1.59)	1.29 (0.73-2.27)	1.42 (0.73-2.75)
IDK	0.55 (0.23-1.29)	0.74 (0.35-1.58)	0.22 (0.04-1.07)	2.48 (0.91-6.74)
Past privacy breach				
Not resolved or IDK	Reference	Reference	Reference	Reference
Resolved	1.15 (0.38-3.43)	1.24 (0.49-3.11)	2.23 (0.68-7.32)	0.57 (0.15-2.13)
No breaches	0.91 (0.36-2.31)	0.91 (0.41-2.01)	1.30 (0.45-3.79)	0.82 (0.28-2.34)
Health care privacy experiences				
Poor	Reference	Reference	Reference	Reference
Good	2.56 (1.24-5.29)[c]	1.46 (0.73-2.90)	1.90 (0.67-5.35)	0.30 (0.20-0.70)[c]
IDK	3.75 (0.97-14.51)	1.81 (0.53-6.19)	1.48 (0.20-10.85)	0.31 (0.07-1.44)
Confidentiality of PHI				
Not private	Reference	Reference	Reference	Reference
Private	1.25 (0.55-2.83)	0.88 (0.41-1.87)	1.50 (0.52-4.39)	1.40 (0.48-4.14)
IDK	0.79 (0.26-2.38)	1.52 (0.54-4.33)	2.14 (0.49-9.32)	2.11 (0.54-8.28)

[a]HCP: health care provider.

[b]DHSP: digital health service provider.

[c]Signifies a significant association when compared with the reference group.

[d]IDK: I do not know.

[e]PNA: prefer not to answer.

[f]PHI: personal health information.

RQ 2: Do Information Needs Differ Among Individuals When Making an Informed Decision to Share Their Health Data With Different Individuals or Entities?

Overall, 81.8% (586/716) of potential users considered sharing their data in both scenarios (ie, potential users who did not select *I do not intend on sharing information with them*). In scenario 1, 89.3% (639/716) of potential users considered granting access to their friends and families and required an average of 5.64 (SD 2.68) of the 9 presented information types to make that decision. In scenario 2, 85.2% (610/716) of potential users considered granting commercial service providers access to

their data and required an average of 5.54 (SD 2.85) of the 9 presented information types to make that decision.

There was no significant difference in the average number of information types required between the 2 scenarios (*P*>.05) for potential users who considered sharing in both scenarios (586/716, 81.8%). On the basis of the frequency of selection by this subset of potential users, the ranking of the types of information differed in the 2 scenarios (Figure 1); however, there was only a significant difference in frequency for 5 of the information types (*P*<.05). Information about accessible data types, restricting access, and revoking was selected more frequently in scenario 1. Information about potential risks and filing complaints was selected more frequently in scenario 2.

Figure 1. Differences in information needs required to support decisions on data sharing with friends and family and commercial service providers (ranked by frequency selected; n=586). PHI: personal health information.
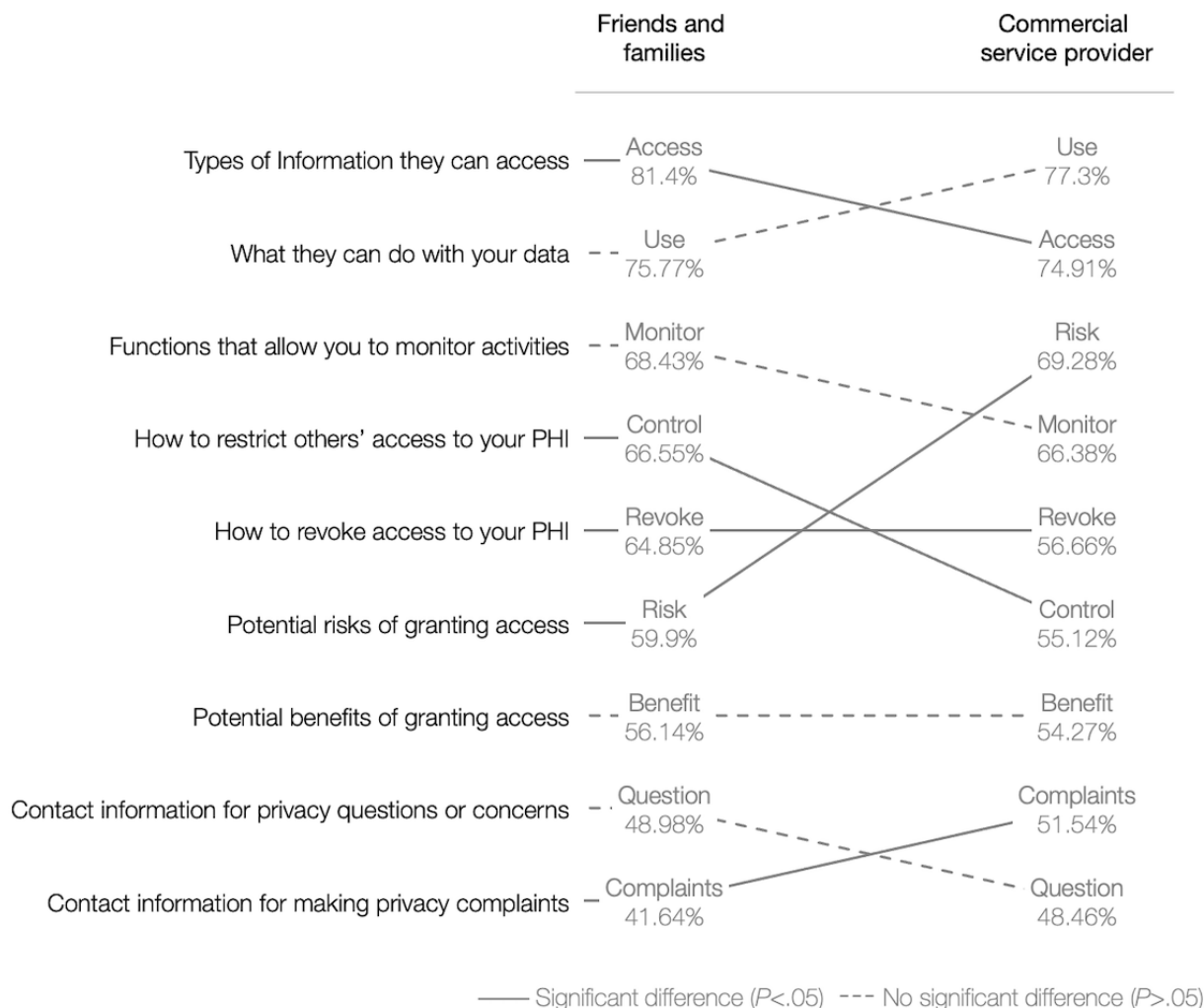
Discussion

Principal Findings

As society becomes increasingly interconnected, there is a corresponding patient anticipation that their PHI and digital health data can be centrally accessed through innovations such as patient access channels, all with the belief that they will make life better [5]. A core requirement critical to the adoption of these patient access channels is a consent management system, as almost all potential users value the ability to control who can access their data. This exploratory study generated some insights to consider when implementing a consent management system. First, there may be acceptance of a believed, broad *all or none* access control model by data source, as 53.8% (385/716) of potential users believed it was adequate for their needs, and 17% (122/716) were unsure. Second, the willingness to provide others with default access to PHI and data varied depending on the recipient. Finally, potential users required an average of approximately 6 key types of information to provide informed decisions regarding data sharing; however, the required types of information varied depending on the recipient. The 3 insights are discussed in detail in the following sections.

Data Control and Consent Management Practices

Given the complexity of implementing interoperable access control in Canada [23], a broad *all or none* access control at the level of the data source may be the only option in the interim for patient access channels, especially as new data sources continuously emerge [8]. If implemented within a context of a *trust framework* in this scenario, there may be an acceptance of broad access control as over half of the potential users believed it was adequate for their needs. This finding echoes that of Grando et al [25], where broad access control was adequate for 58% of their study participants; moreover, their study was set in the context of behavioral health—an area where PHI is often perceived as more sensitive. Similarly, and surprisingly, user perceptions of the sensitivity of PHI and digital health data were not associated with adequacy, especially as data sensitivity is commonly associated with wanting greater degrees of access control because of privacy concerns [26,27]. A possible explanation is that the sample had a high level of trust in the confidentiality of their PHI and had positive dispositions about their web-based and health care privacy. This is consistent with an emerging set of evidence showing that positive perceptions of health care, trust in health care providers, and positive past

privacy experiences may result in individuals having favorable views on sharing data [28-34]. Although these studies are contrary to prior findings of patients wanting more granular control options [25,27,35], their hypothetical and exploratory nature is subject to the privacy paradox [36]—the disconnect between intentions based on privacy concerns and actual behaviors. For instance, Schwartz et al [37] provided 108 patients with the ability to restrict access to their sensitive EHR data and found that 57% provided access to all listed providers and all PHI in their EHR, and 8.6% limited access by data type to specific providers. A significant minority of participants (43%) limited access to at least one provider.

Approximately one-fifth of the potential users did not know whether broad access control would be adequate, highlighting the need to better support their decision-making. The technical aspects of sharing data may be complex and may require greater literacy to appreciate the impact of broad access control [23]. This may explain why digital health use was associated with a 109% increase in the likelihood of knowing whether it is adequate. Familiarity and experience with digital health may provide individuals with heuristics to make decisions [34]. Studies show that broad access control and consent models may be acceptable when there is transparency [28,38,39] and assurance in oversight [40] regarding how the data are used. Biobank studies have shown that there are no significant differences in the willingness to share data between various consent scenarios when participants are provided with specific information on the data that are being used [38] or if there is assurance that a governing body provides oversight on how data are being used [40]. These findings can be applied to the digital health context, as a recent survey found that 80% of Canadians are willing to share their anonymized health information as long as the privacy and security of their PHI are assured [41].

Income and region were the only demographics found to be associated with adequacy perspectives on broad access control and knowing whether broad access control was adequate. Although the association between privacy attitudes and income echoes some privacy studies in health informatics, there have often been conflicting results across studies [33]. Historically, privacy research has focused on demographic variables as predictors of privacy attitudes and behaviors; however, collective evidence signals that individual demographic variables play a minor role and provide limited insight into understanding a phenomenon [33,42]. These findings are intended to inform further explorations to support implementation decisions. For instance, there may be value in understanding the underlying factors associated with those with high incomes that support their views of inadequacy and why they are more likely to know whether broad access control reflects their needs. In terms of region, health care in Canada is administered at the provincial level, where there are variations in legislation, policies, and digital health initiatives. Only a few provinces in Canada have a centralized patient portal, and the Prairie provinces of Alberta and Saskatchewan were launching theirs at the time of this study [43-45]. Understanding how these initiatives may have affected attitudes on broad access control adequacy may inform strategies on how to improve public favorability toward broad access control.

Potential users were most willing to grant default access to their health care providers, especially those with positive health care privacy experiences. Their willingness decreased as the data recipient was further removed from the point of care. Patients generally trust their physicians and those in their circle of care to keep their data confidential; however, this trust to maintain confidentiality diminishes as the data recipient is further away from those providing care (eg, health department, researchers, and corporations) [25,27,35,46]. Canadians are generally comfortable with the sharing of their PHI through EHRs with other health care providers as they believe timely and easy access to PHI is necessary for high-quality care [47], highlighting the role of contextual relevance and issue involvement [48] in data-sharing behaviors. This assertion is supported by the finding that those with prior digital health use were 66% to 117% more likely to grant default access (entity dependent) than those who did not use digital health tools. These users may have a greater stake in using digital health tools, familiarity, and perceived benefits of sharing digital health data [28-34]. Contextual relevance also mattered for users with higher perceived sensitivity of PHI, as 126% were more likely to not give anyone default access and were less likely to grant default access to family, friends, and other supporters. These individuals may not be comfortable with default access and may want more control over how new information is shared. Sharing may depend on the purpose and whether it is a necessity; moreover, these individuals may want more control over how certain information is disclosed to close social associates as it may affect their relationships. They may want to share about it in person rather than have others find it out by default through technology [34].

The value of data sharing with digital health service providers may not be as clear as there is limited trust in service providers, especially commercial vendors [19,47]. In this study, one-fifth of potential users were willing to grant default access to service providers, of whom users with prior digital health experience and caregivers were more likely to share their data. As discussed earlier, these users may have greater perceived benefits of granting default access to data to service providers [28-34]. For caregivers, sharing data for this population may be perceived to improve the tasks and stressors associated with their caregiving roles through the development of better or improved digital health services [49]. Further understanding the rationales of those trusting and skeptical of commercial service providers will be a necessity as these providers are a growing contributor to the number and types of services provided and an important source of data for patient access channels. This understanding can inform the permitted uses outlined in the trust framework and enable informed and meaningful consent.

Information Needs

This study builds upon the Meaningful Consent Guidelines for use in the digital health context. The guidelines recommend that digital vendors emphasize four key information elements: what is collected, who has access, purpose of data collection, and potential risks. However, this study found that potential users may need 5 to 6 emphasized information elements. The additional elements of emphasis include information on monitoring access, restricting access, and revoking consent.

The findings also suggest that there is a need to tailor the order of emphasized elements as they will vary depending on who is accessing the data.

This study also highlights the importance of patient engagement in ensuring that the design of consent is based on user needs rather than assumptions. For instance, presenting this consent information in clear, concise, and plain language has been advocated but seldom practiced; however, implementing this assumption is only a part of the solution. A study found that an easier to read, concise consent form neither hindered nor improved comprehension or satisfaction with the consent process among their participants [50]. In contrast, providing users with ways of customizing their experiences and consuming information is more effective [51]. User experience is an overlooked aspect that should be considered when implementing informed and meaningful consent [23,52]. To empower users to make informed choices, meaningful consent for patient access channels should be iteratively co-designed with its users to ensure that they meet their needs rather than their assumed wants [53].

Limitations

This study provides preliminary insights to support future patient engagement in co-designing a consent management system and meaningful consent. However, these exploratory findings are not intended to be generalizable as there are limitations to consider. This study is a secondary analysis of a cross-sectional survey, providing a snapshot of a time point where perspectives may vary over time. This study relied on a series of vignettes to preface the questions. Multiple rounds of revisions were made with Canada Health Infoway's communications department and the market research firm to improve clarity of complex concepts (eg, privacy, consent, and data sharing). Prompts with these concepts include languages with high readability scores, which may have influenced some responses, especially those with lower digital literacy skills [54].

There are also inherent limitations to data collection through a survey panel as it only includes people who participate in the web panels managed by the company and relies on the self-selection of participants. The web-based nature of the survey

may have excluded the perspectives of individuals with limited internet access. However, approximately 94% of Canadian households currently have access to the internet [55]. The purposive sampling strategy limits generalizability to the broader Canadian population as recruitment focused on frequent users of health care and excluded the Canadian territories (ie, early adopters of a patient access channel) [56,57]. The identified users in this study may be more engaged and experienced with digital health tools, thereby perceiving greater benefits and a greater willingness to share their data. The low response rate should also be considered as it may limit the diversity and nuance of perspectives because of the information lost through the combination or omission of demographics and participant characteristics for data analysis (eg, individuals who are transgender and other identifying genders). Future public and stakeholder engagement activities will require a greater in-depth investigation in co-designing consent management for patient access channels. Recognizing the ethical transgressions in trust in health care and research of marginalized and vulnerable communities [58], future research must include more diversity in perspectives to understand how to equitably strengthen meaningful consent and consent management practices.

Conclusions

Providing patients with the ability to manage their consent and control access to their PHI is valued by potential users of a patient access channel. Following the Office of the Privacy Commissioner of Canada's Meaningful Consent Guidelines, future work should continue to *consider the consumer's perspective* by involving them throughout the development and implementation processes [18]. Given technological limitations, future public engagement should investigate what makes broad access control acceptable and how to communicate its implications meaningfully and transparently. Future research should also focus on understanding user requirements for consent to further adapt the Meaningful Consent Guidelines for the digital health context. Understanding how to foster patient trust and how to empower them to feel confident in their data-sharing decisions is necessary for the success of patient access channels and the realization of the transformative potential of the evolving digital health ecosystem.

Conflicts of Interest

SW and ACL (affiliated with Canada Health Infoway) were members of the research team and were involved in the study design, interpretation and manuscript development.

Multimedia Appendix 1

Consent scenarios and vignettes.
[DOCX File , 19 KB - medinform_v10i3e30986_app1.docx]

Multimedia Appendix 2
Covariate definitions and outcome variables for logistic regression.
[DOCX File , 18 KB - medinform_v10i3e30986_app2.docx]

Multimedia Appendix 3
Full survey questionnaire.
[DOCX File , 141 KB - medinform_v10i3e30986_app3.docx]

References

1. Connecting patients for better health. Canada Health Infoway. 2016. URL: https://www.infoway-inforoute.ca/en/component/edocman/resources/3152-connecting-patients-for-better-health-2016 [accessed 2021-11-30]

2. Connecting patients for better health: 2018. Canada Health Infoway. 2018. URL: https://www.infoway-inforoute.ca/en/component/edocman/resources/reports/benefits-evaluation/3564-connecting-patients-for-better-health-2018 [accessed 2021-11-30]

3. Symons JD, Ashrafian H, Dunscombe R, Darzi A. From EHR to PHR: let's get the record straight. BMJ Open 2019;9(9):e029582 [FREE Full text] [doi: 10.1136/bmjopen-2019-029582] [Medline: 31537566]

4. Sterud B. Practitioner Application: the challenges in personal health record adoption. J Healthc Manag 2019;64(2):109-110. [doi: 10.1097/JHM-D-19-00010] [Medline: 30845059]

5. The future of connected health care: reporting Canadians' perspective on the health care system. Canadian Medical Association. 2019. URL: https://www.cma.ca/sites/default/files/pdf/Media-Releases/The-Future-of-Connected-Healthcare-e.pdf [accessed 2021-11-30]

6. Smith HJ, Dinev T, Xu H. Information privacy research: an interdisciplinary review. MIS Q 2011;35(4):989-1015. [doi: 10.2307/41409970]

7. Health data privacy and access across Canada. Canada Health Infoway. 2021. URL: https://www.infoway-inforoute.ca/en/patients-families-caregivers/digital-health-learning-program/get-familiar-with-health-data [accessed 2021-11-30]

8. Asghar MR, Lee T, Baig MM, Ullah E, Russello G, Dobbie G. A review of privacy and consent management in healthcare: a focus on emerging data sources. In: IEEE 13th International Conference on e-Science. 2017 Presented at: e-Science '17; October 24-27, 2017; Auckland, New Zealand p. 518-522. [doi: 10.1109/escience.2017.84]

9. Zazaza L, Venter HS, Sibiya G. The current state of electronic consent systems in e-health for privacy preservation. In: Proceedings of the 17th International Information Security Conference. 2018 Presented at: ISSA '18; August 15-16, 2018; Pretoria, South Africa p. 76-88. [doi: 10.1007/978-3-030-11407-7_6]

10. Canada's digital charter: trust in a digital world. Government of Canada. 2020. URL: https://www.ic.gc.ca/eic/site/062.nsf/eng/h_00108.html [accessed 2020-12-07]

11. Martinez-Martin N, Kreitmair K. Ethical issues for direct-to-consumer digital psychotherapy apps: addressing accountability, data protection, and consent. JMIR Ment Health 2018;5(2):e32 [FREE Full text] [doi: 10.2196/mental.9423] [Medline: 29685865]

12. Sharon T. When digital health meets digital capitalism, how many common goods are at stake? Big Data Soc 2018;5(2):205395171881903. [doi: 10.1177/2053951718819032]

13. World Economic Forum. 2020. URL: https://www.weforum.org/reports/redesigning-data-privacy-reimagining-notice-consent-for-humantechnology-interaction [accessed 2021-11-30]

14. Litman-Navarro K. We read 150 privacy policies. They were an incomprehensible disaster. The New York Times. 2019. URL: https://www.nytimes.com/interactive/2019/06/12/opinion/facebook-google-privacy-policies.html [accessed 2021-11-30]

15. Sunyaev A, Dehling T, Taylor PL, Mandl KD. Availability and quality of mobile health app privacy policies. J Am Med Inform Assoc 2015;22(e1):e28-e33. [doi: 10.1136/amiajnl-2013-002605] [Medline: 25147247]

16. Armstrong S. Data, data everywhere: the challenges of personalised medicine. BMJ 2017;359:j4546. [doi: 10.1136/bmj.j4546] [Medline: 29021195]

17. Nebeker C, Torous J, Bartlett Ellis RJ. Building the case for actionable ethics in digital health research supported by artificial intelligence. BMC Med 2019;17(1):137 [FREE Full text] [doi: 10.1186/s12916-019-1377-7] [Medline: 31311535]

18. Guidelines for obtaining meaningful consent. Office of the Privacy Commissioner of Canada. 2018. URL: https://www.priv.gc.ca/en/privacy-topics/collecting-personal-information/consent/gl_omc_201805/ [accessed 2021-11-30]

19. Torous J, Roberts LW. Needed innovation in digital health and smartphone applications for mental health: transparency and trust. JAMA Psychiatry 2017;74(5):437-438. [doi: 10.1001/jamapsychiatry.2017.0262] [Medline: 28384700]

20. Greenhalgh T, Wherton J, Papoutsi C, Lynch J, Hughes G, A'Court C, et al. Beyond adoption: a new framework for theorizing and evaluating nonadoption, abandonment, and challenges to the scale-up, spread, and sustainability of health and care technologies. J Med Internet Res 2017;19(11):e367 [FREE Full text] [doi: 10.2196/jmir.8775] [Medline: 29092808]

21. Duggal R, Brindle I, Bagenal J. Digital healthcare: regulating the revolution. BMJ 2018;360:k6. [doi: 10.1136/bmj.k6] [Medline: 29335296]

22. Shen N, Strauss J, Silver M, Carter-Langford A, Wiljer D. The eHealth trust model: a patient privacy research framework. Stud Health Technol Inform 2019;257:382-387. [Medline: 30741227]

23. Shen N, Kassam I, Ilkina D, Wickham S, Carter-Langford A. Meaningful digital consent in Canada: recommendations from pan-Canadian consent management workshops. Healthc Q 2022;24(4):40-47. [doi: 10.12927/hcq.2022.26712] [Medline: 35216648]

24. DiStefano C, Shi D, Morgan GB. Collapsing categories is often more advantageous than modeling sparse data: investigations in the CFA framework. Struct Equ Model 2021;28(2):237-249. [doi: 10.1080/10705511.2020.1803073]

25. Grando MA, Murcko A, Mahankali S, Saks M, Zent M, Chern D, et al. A study to elicit behavioral health patients' and providers' opinions on health records consent. J Law Med Ethics 2017;45(2):238-259 [FREE Full text] [doi: 10.1177/1073110517720653] [Medline: 30976154]

26. Serrano KJ, Yu M, Riley WT, Patel V, Hughes P, Marchesini K, et al. Willingness to exchange health information via mobile devices: findings from a population-based survey. Ann Fam Med 2016;14(1):34-40 [FREE Full text] [doi: 10.1370/afm.1888] [Medline: 26755781]

27. Caine K, Hanania R. Patients want granular privacy control over health information in electronic medical records. J Am Med Inform Assoc 2013;20(1):7-15 [FREE Full text] [doi: 10.1136/amiajnl-2012-001023] [Medline: 23184192]

28. Esmaeilzadeh P. The impacts of the perceived transparency of privacy policies and trust in providers for building trust in health information exchange: empirical study. JMIR Med Inform 2019;7(4):e14050 [FREE Full text] [doi: 10.2196/14050] [Medline: 31769757]

29. Esmaeilzadeh P, Mirzaei T. Comparison of consumers' perspectives on different health information exchange (HIE) mechanisms: an experimental study. Int J Med Inform 2018;119:1-7. [doi: 10.1016/j.ijmedinf.2018.08.007] [Medline: 30342677]

30. Li T, Slee T. The effects of information privacy concerns on digitizing personal health records. J Assn Inf Sci Tec 2014;65(8):1541-1554. [doi: 10.1002/asi.23068]

31. Maiorana A, Steward WT, Koester KA, Pearson C, Shade SB, Chakravarty D, et al. Trust, confidentiality, and the acceptability of sharing HIV-related patient data: lessons learned from a mixed methods study about Health Information Exchanges. Implement Sci 2012;7:34 [FREE Full text] [doi: 10.1186/1748-5908-7-34] [Medline: 22515736]

32. Walker DM, Johnson T, Ford EW, Huerta TR. Trust me, I'm a doctor: examining changes in how privacy concerns affect patient withholding behavior. J Med Internet Res 2017;19(1):e2 [FREE Full text] [doi: 10.2196/jmir.6296] [Medline: 28052843]

33. Shen N, Bernier T, Sequeira L, Strauss J, Silver MP, Carter-Langford A, et al. Understanding the patient privacy perspective on health information exchange: a systematic review. Int J Med Inform 2019;125:1-12. [doi: 10.1016/j.ijmedinf.2019.01.014] [Medline: 30914173]

34. Shen N, Sequeira L, Silver MP, Carter-Langford A, Strauss J, Wiljer D. Patient privacy perspectives on health information exchange in a mental health context: qualitative study. JMIR Ment Health 2019;6(11):e13306 [FREE Full text] [doi: 10.2196/13306] [Medline: 31719029]

35. Soni H, Grando A, Aliste MP, Murcko A, Todd M, Mukundan M, et al. Perceptions and preferences about granular data sharing and privacy of behavioral health patients. Stud Health Technol Inform 2019;264:1361-1365. [doi: 10.3233/SHTI190449] [Medline: 31438148]

36. Acquisti A, Brandimarte L, Loewenstein G. Privacy and human behavior in the age of information. Science 2015;347(6221):509-514. [doi: 10.1126/science.aaa1465] [Medline: 25635091]

37. Schwartz PH, Caine K, Alpert SA, Meslin EM, Carroll AE, Tierney WM. Patient preferences in controlling access to their electronic health records: a prospective cohort study in primary care. J Gen Intern Med 2015;30 Suppl 1:S25-S30 [FREE Full text] [doi: 10.1007/s11606-014-3054-z] [Medline: 25480721]

38. Kaufman DJ, Baker R, Milner LC, Devaney S, Hudson KL. A survey of U.S adults' opinions about conduct of a nationwide precision medicine initiative® cohort study of genes and environment. PLoS One 2016;11(8):e0160461 [FREE Full text] [doi: 10.1371/journal.pone.0160461] [Medline: 27532667]

39. Esmaeilzadeh P. The effect of the privacy policy of Health Information Exchange (HIE) on patients' information disclosure intention. Comput Secur 2020;95:101819. [doi: 10.1016/j.cose.2020.101819]

40. Sanderson SC, Brothers KB, Mercaldo ND, Clayton EW, Antommaria AH, Aufox SA, et al. Public attitudes toward consent and data sharing in biobank research: a large multi-site experimental survey in the US. Am J Hum Genet 2017;100(3):414-427 [FREE Full text] [doi: 10.1016/j.ajhg.2017.01.021] [Medline: 28190457]

41. Consulting Canadians on the future of their health system: a health dialogue. Canada Health Infoway. 2020. URL: https://www.infoway-inforoute.ca/en/component/edocman/resources/reports/3850-a-healthy-dialogue-executive-summary [accessed 2021-11-30]

42. Gerber N, Gerber P, Volkamer M. Explaining the privacy paradox: a systematic review of literature investigating privacy attitude and behavior. Comput Secur 2018;77:226-261. [doi: 10.1016/j.cose.2018.04.002]

43. Avdagovska M, Stafinski T, Ballermann M, Menon D, Olson K, Paul P. Tracing the decisions that shaped the development of MyChart, an electronic patient portal in Alberta, Canada: historical research study. J Med Internet Res 2020;22(5):e17505 [FREE Full text] [doi: 10.2196/17505] [Medline: 32452811]

44. Basky G. Some provinces still delay access to health records via patient portals. CMAJ 2019;191(48):E1341 [FREE Full text] [doi: 10.1503/cmaj.1095829] [Medline: 31791973]

45. New website allows Saskatchewan residents to access their personal health information anywhere, anytime. Government of Saskatchewan. 2019. URL: https://www.saskatchewan.ca/government/news-and-media/2019/october/08/ehealth-website [accessed 2021-11-30]

46. Hunter IM, Whiddett RJ, Norris AC, McDonald BW, Waldon JA. New Zealanders' attitudes towards access to their electronic health records: preliminary results from a national study using vignettes. Health Informatics J 2009;15(3):212-228 [FREE Full text] [doi: 10.1177/1460458209337435] [Medline: 19713396]

47. Canadian digital health survey: what Canadians think. Canada Health Infoway. 2020. URL: https://www.infoway-inforoute.ca/en/component/edocman/resources/reports/benefits-evaluation/3856-canadian-digital-health-survey-what-canadians-think [accessed 2021-11-30]

48. Abdelhamid M, Gaia J, Sanders GL. Putting the focus back on the patient: how privacy concerns affect personal health information sharing intentions. J Med Internet Res 2017;19(9):e169 [FREE Full text] [doi: 10.2196/jmir.6877] [Medline: 28903895]

49. Lindeman DA, Kim KK, Gladstone C, Apesoa-Varano E. Technology and caregiving: emerging interventions and directions for research. Gerontologist 2020;60(Suppl 1):S41-S49 [FREE Full text] [doi: 10.1093/geront/gnz178] [Medline: 32057082]

50. Grady C, Touloumi G, Walker AS, Smolskis M, Sharma S, Babiker AG, INSIGHT START Informed Consent Substudy Group. A randomized trial comparing concise and standard consent forms in the START trial. PLoS One 2017;12(4):e0172607 [FREE Full text] [doi: 10.1371/journal.pone.0172607] [Medline: 28445471]

51. Beskow LM, Friedman JY, Hardy NC, Lin L, Weinfurt KP. Developing a simplified consent form for biobanking. PLoS One 2010;5(10):e13302 [FREE Full text] [doi: 10.1371/journal.pone.0013302] [Medline: 20949049]

52. Is "Meaningful Consent" a contradiction in terms?: three design jams seek the answer. Office of the Privacy Commissioner of Canada. 2021. URL: https://www.priv.gc.ca/en/opc-actions-and-decisions/research/funding-for-privacy-research-and-knowledge-translation/real-results/rr-index/jam-intro/ [accessed 2021-11-30]

53. Adam MB, Minyenya-Njuguna J, Karuri Kamiru W, Mbugua S, Makobu NW, Donelson AJ. Implementation research and human-centred design: how theory driven human-centred design can sustain trust in complex health systems, support measurement and drive sustained community health volunteer engagement. Health Policy Plan 2020;35:ii150-ii162 [FREE Full text] [doi: 10.1093/heapol/czaa129] [Medline: 33156944]

54. McKnight DH, Choudhury V, Kacmar C. Developing and validating trust measures for e-commerce: an integrative typology. Inf Syst Res 2002;13(3):334-359. [doi: 10.1287/isre.13.3.334.81]

55. Access to the internet in Canada, 2020. Statistics Canada. 2020. URL: https://www150.statcan.gc.ca/n1/daily-quotidien/210531/dq210531d-eng.htm [accessed 2021-11-30]

56. Mák G, Smith Fowler HS, Leaver C, Hagens S, Zelmer J. The effects of web-based patient access to laboratory results in British Columbia: a patient survey on comprehension and anxiety. J Med Internet Res 2015;17(8):e191 [FREE Full text] [doi: 10.2196/jmir.4350] [Medline: 26242801]

57. Leonard KJ, Casselman M, Wiljer D. Who will demand access to their personal health record? A focus on the users of health services and what they want. Healthc Q 2008;11(1):92-96 [FREE Full text] [doi: 10.12927/hcq.2008.19503] [Medline: 18326386]

58. O'Sullivan L, Crowley R, McAuliffe É, Doran P. Contributory factors to the evolution of the concept and practice of informed consent in clinical research: a narrative review. Contemp Clin Trials Commun 2020;19:100634 [FREE Full text] [doi: 10.1016/j.conctc.2020.100634] [Medline: 33024880]

Abbreviations

CIHR: Canadian Institutes of Health Research
EHR: electronic health record
OR: odds ratio
PHI: personal health information
RQ: research question

XSL•FO
RenderX

Original Paper

A Comparison of Census and Cohort Sampling Models for the Longitudinal Collection of User-Reported Data in the Maternity Care Pathway: Mixed Methods Study

Kendall Jamieson Gilmore[1], MSc; Manila Bonciani[1], PhD; Milena Vainieri[1], PhD

Management and Healthcare Laboratory, Department of Economics and Management in the era of Data Science, Institute of Management, Sant'Anna Scuola Superiore, Pisa, Italy

Corresponding Author:
Kendall Jamieson Gilmore, MSc
Management and Healthcare Laboratory
Department of Economics and Management in the era of Data Science, Institute of Management
Sant'Anna Scuola Superiore
33 Piazza Martiri della Libertà
Pisa, 56127
Italy
Phone: 39 050 883111
Email: k.jamiesongilmore@santannapisa.it

Abstract

Background: Typical measures of maternity performance remain focused on the technical elements of birth, especially pathological elements, with insufficient measurement of nontechnical measures and those collected pre- and postpartum. New technologies allow for patient-reported outcome measures (PROMs) and patient-reported experience measures (PREMs) to be collected from large samples at multiple time points, which can be considered alongside existing administrative sources; however, such models are not widely implemented or evaluated. Since 2018, a longitudinal, personalized, and integrated user-reported data collection process for the maternal care pathway has been used in Tuscany, Italy. This model has been through two methodological iterations.

Objective: The aim of this study was to compare and contrast two sampling models of longitudinal user-reported data for the maternity care pathway, exploring factors influencing participation, cost, and suitability of the models for different stakeholders.

Methods: Data were collected by two modes: (1) "cohort" recruitment at the birth hospital of a predetermined sample size and (2) continuous, ongoing "census" recruitment of women at the first midwife appointment. Surveys were used to collect experiential and outcome data related to existing services. Women were included who passed 12 months after initial enrollment, meaning that they either received the surveys issued after that interval or dropped out in the intervening period. Data were collected from women in Tuscany, Italy, between September 2018 and July 2020. The total sample included 7784 individuals with 38,656 observations. The two models of longitudinal collection of user-reported data were analyzed using descriptive statistics, survival analysis, cost comparison, and a qualitative review.

Results: Cohort sampling provided lower initial participation than census sampling, although very high subsequent response rates (87%) were obtained 1 year after enrollment. Census sampling had higher initial participation, but greater dropout (up to 45% at 1 year). Both models showed high response rates for online surveys. There were nonproportional dropout hazards over time. There were higher rates of dropout for women with foreign nationality (hazard ratio [HR] 1.88, $P<.001$), and lower rates of dropout for those who had a higher level of education (HR 0.77 and 0.61 for women completing high school and college, respectively; $P<.001$), were employed (HR 0.87, $P=.01$), in a relationship (HR 0.84, $P=.04$), and with previous pregnancies (HR 0.86, $P=.002$). The census model was initially more expensive, albeit with lower repeat costs and could become cheaper if repeated more than six times.

Conclusions: The digital collection of user-reported data enables high response rates to targeted surveys in the maternity care pathway. The point at which pregnant women or mothers are recruited is relevant for response rates and sample bias. The census model of continuous enrollment and real-time data availability offers a wider set of potential benefits, but at an initially higher cost and with the requirement for more substantial data translation and managerial capacity to make use of such data.

XSL•FO

RenderX

Introduction

Most health care performance data are derived from administrative sources, which can be used to measure the more technical aspects or process measures of maternity care provided to women. However, such data can only capture some dimensions of the quality of care and do not address important features such as patient preferences or overall well-being. These data are also limited since they are collected through interactions with health care providers, do not usually relate to community-based settings, and cannot provide insights outside of formal interactions with a subset of health services, thereby potentially neglecting the contexts in which the real value of care delivered becomes apparent [1-3].

More recently, studies highlighting the importance of priority setting, use of management models, incentives, and other similar efforts have shown that data related to patient-centered measurement may prove to be more useful [4-6]. This includes assessments of patients' preferences for care, experiences with services, and a range of disease-specific and general health and well-being–related markers. These latter two domains are typically collected through validated tools such as patient-reported experience measures (PREMs) and patient-reported outcome measures (PROMs), respectively [7,8], which can provide responsive and reliable measures of outcomes and experiences [9].

Although they are mainly used in measuring experiences pertaining to or outcomes resulting from acute health care, PREMs and PROMs can be used as longer-term, longitudinal measures. Such measures may include outcomes not tied to specific interactions with health care professionals, such as the case of chronic conditions that are primarily managed by patients themselves. A wide set of characteristics can be measured in this way, including those more relevant to patients' quality of life than clinical or administrative markers [10]. Technological advances now allow for the systematic collection and analysis of large amounts of such data. Survey data can be collected from large samples at multiple time points; where such models are implemented, the definition of routinely collected data is effectively broadened to include PROMs and PREMs. Although technically feasible, such models have not yet been widely implemented [10].

The maternity care pathway is well-suited to such technology-enabled collection of PROMs and PREMs at scale, since typical performance indicators remain focused on the technical elements of birth, especially pathological elements. However, there is insufficient collection and use of person-centered indicators and of quality measures along the pathway [11], despite evidence that women's experiences of childbirth go far beyond labor, and that social and psychological aspects of care are important for women [12,13]. Indeed, World Health Organization recommendations underline the importance of woman-centered care to optimize the experience of pregnancy, labor, and childbirth for women and babies through a holistic, human rights–based approach, promoting continuity of care along the pathway [14,15]. Existing efforts to address this information gap include birth cohort studies using online surveys advertised to women through posters and leaflets [16], online surveys advertised through social media [12], and cross-sectional national postal surveys of randomly selected women [17].

For several years, the Tuscan maternal care pathway performance evaluation has adopted a pathway perspective, framed around the person, with inclusion of PREMs collected through periodic patient surveys [18]. More recently, the model was developed through use of digital technology to include both PREMs and PROMs collected at multiple points in time. This model has been implemented in two methodological versions: a cohort-sampling model and a census-sampling model. Both are fully digital, with pregnant women or new mothers enrolled by health professionals and subsequently contacted by SMS text message or email containing survey links.

Previous evaluations of maternity survey data–collection models focused on comparing alternative models based around cross-sectional postal surveys [17]. This study provides insight into two models of longitudinal, personalized, and integrated data collection, the nature of which enable use of analytical methods that have not, to our knowledge, previously been applied in this context. Comparing the performance of the models is useful from several aspects. First, this serves to describe and evaluate each model in isolation, as both offer new features and insights compared to typical maternity data. Second, there remain notable differences in the capabilities and costs between the models.

Thus, the primary research question for this study was: what are the conditions in which a census model is preferable to a cohort-sampling model of longitudinal data collection in the maternity pathway? To answer to this question, we explored (1) whether there are differences in survey participation rates between the two data-collection models; (2) which characteristics of the samples affect participation in the survey; (3) the costs of census and cohort sampling methods; and (4) the strengths and weakness of census and cohort sampling methods, considering the usefulness of the two data collection models to different stakeholders.

Methods

Study Design

We used a mixed methods approach to compare several dimensions of census and cohort sampling models of longitudinal data collection along the maternity pathway in Tuscany, Italy. For the quantitative component of the research,

we applied survival analysis of survey data, along with identification and comparison of costs, whereas for the qualitative component, we compared the two models with respect to different audiences and purposes.

Data Source: Longitudinal Data Collection Models

Longitudinal user-reported data collection was carried out from September 2018 to March 2019, using a cohort sampling model. In this model, a predetermined number of women are recruited at maternity hospitals after birth. The sample size was calculated to be representative at the hospital and district levels, considering a 95% confidence level with a 7%-9% 95% CI and 20% follow-up loss. This resulted in a required sample of 3672 women (which was lower than the number ultimately recruited). Exclusion criteria were being a resident outside of Tuscany, preterm delivery (<37 weeks of gestation) or low birth weight (<2500 grams) newborn, or hospitalized in neonatal intensive care. Enrollment was led by midwives in birth hospitals, who were asked to invite every woman (without exclusion criteria) that they supported in birth until the target was reached, after which they were advised to stop. All 24 birth hospitals in Tuscany participated in the recruitment. Each woman was asked to complete six surveys, covering the period from delivery to 12 months later. These surveys were completed after discharge and at 1, 3, 5, 6, and 12 months after childbirth. The follow-up was therefore complete in March 2020.

The second model of longitudinal data collection, which has been in continuous operation since March 2019, uses a census sampling approach. All pregnant women residing in Tuscany who withdrew their pregnancy booklet from a family care center were eligible, with women recruited continuously in these facilities. All 117 family care centers, as the first contact point in the maternal care pathway, participated in the data collection.

Midwives enroll all consenting women on an ongoing basis, utilizing the integration between the survey system and the regional information system recording data for pregnancy booklet delivery. Each woman is asked to respond to eight surveys, with data collected at the following points: receipt of the pregnancy booklet; in the second trimester of pregnancy; third trimester of pregnancy; at expected childbirth; and at 1, 3, 6, and 12 months after childbirth. As of the end of 2019, 10,821 women were involved in the survey and answered the first questionnaire.

All surveys are available in Italian and in the seven most commonly spoken foreign languages in Tuscany: English, French, Spanish, Albanian, Arabic, Romanian, and Chinese.

Both models are underpinned by a digitized process. First, the approach is explained to expectant or new mothers, including information about privacy and data uses. Women who consent to join are enrolled in the system, after which surveys are automatically administered according to the stage of the maternity pathway. The questionnaires are personalized according to the gestational period of pregnancy or age of the newborn (only the latter for the cohort sampling model) and include both PROM and PREM items. Each survey addresses topics that are important for pregnant women and new mothers (Figure 1). Unique links to the online surveys are shared by email or SMS, to be filled in by women in a time and place of their choosing using any web-enabled device. Up to three reminders are sent for each survey. Responses are automatically collated and hosted in a secure web platform. Results from the cohort model are presented in a research report at the end of the follow-up period, whereas the census model uses real-time return of data to managers and professionals through a web platform.

Figure 1. Characteristics of the census and cohort sampling methods. The survey abbreviations in the cohort sampling model refer to the month at which the survey was issued postbirth (ie, T3 in the cohort model was issued at 3 months postbirth). The time points for the census model represent the trimesters in pregnancy (gravidanza in Italian, "g") and months postpartum ("p"). The pregnancy time points represent the trimesters (ie, T3g is the third trimester) and postbirth points represent months after birth, as in the cohort model (ie, T6p is issued 6 months postbirth). The estimated completion time for each questionnaire is based on the upper and lower limits of items (depending on responses to screener questions) and the type of questions per survey. PREM: patient-reported experience measure; PROM: patient-reported outcome measure; QoL: quality of life.

For this study, we included data for the cohort model from all mothers recruited between September 2018 and March 2019 (3849 women completed the first questionnaire). All women could have reached the final survey 12 months after enrollment (although some may have dropped out earlier).

For the census sampling model, the data extracted also covered women who joined the data collection period sufficiently long ago to have been able to reach the survey issued 12 months after joining the data collection (wave 6, T3p in Figure 1). The information of women who joined the census sampling model less than 12 months before data collection were extracted and excluded from the analysis. The resultant time period of data collection of the six waves from the census model was from March 2019 to July 2020, with 3935 women included in the study. Data were combined into a single pooled database.

Data Analysis

Response and attrition rates within each sample population were calculated with reference to the total number of women enrolled from delivery, in the case of the cohort sampling model, and from the first midwife appointment (delivery of the pregnancy booklet), in the case of the census sampling model. The total period of recruitment and data collection for the cohort model was included. Enrollment rates were calculated for both models, using the numbers of enrolled, responding, and total eligible women.

Our methods for interrogating survey attrition rates for women who elected to join the data collection model were based on the framework proposed by Hochheimer et al [19] for evaluating attrition in web-administered surveys. This includes several steps: visualizing attrition using bar charts and survival-type curves; investigation using generalized linear mixed models

(GLMMs); and further analysis using survival analysis such as Kaplan-Meier curves, log-rank test, and Cox proportional hazards. Each step provides an additional level of granularity, which can be tailored as appropriate to the study circumstance [19]. We followed all steps except for the GLMM investigating sequential questions, since we were interested in comparing the overall survival in the two models and the factors that are explanatory of survival, rather than determining the significance of changes between sequential questions (in this case, survey waves).

Survival analysis was performed in line with the methodology described above along with processes that were additionally considered appropriate (rather than discrete-time modeling approaches) for our data due to the unequally spaced survey waves [20]. Survival was defined as completion of all eligible survey waves. We created a new entry in the pooled database indicating the failure point for individuals dropping out of the data collection before completion. In this way, the failure event was defined as the occasion in which an individual *could have* responded to a survey *but did not* (rather than the last survey to which they *did* respond). Respondents who completed all eligible survey waves were considered censored. Women in the census model who had an abortion were excluded from analyses. The total population for survival analysis, comprising women from both data collection models, was 7784 individuals with 38,656 observations.

The regression model was built through sequential univariate testing of variables and testing for interaction terms, followed by testing the resulting full model. For categorical variables, the log-rank test of equality was used, with Cox proportional hazard regression used for age (the only continuous variable).

The final multivariate model was built using Cox proportional hazards. Covariates included in the final model were age, foreign status (dummy variable), education level (scored from 1-3, with 1=less than high school, 2=high school diploma, and 3=college graduate), employment status (dummy variable), relationship status (dummy variable), and whether the woman had previously had a child (nulliparous, dummy variable).

Qualitative Comparison

Comparative analysis of the strengths and challenges of each model was performed, informed by theory and by discussions with project stakeholders to reflect their perspectives of the usefulness of the two longitudinal data collection models. These aspects are summarized according to methodological factors, managerial factors, and evaluation factors.

Cost Comparison

Comparative cost analysis was performed to illustrate the effect of time and number of cohort samples on relative cost-effectiveness, using available program cost data and estimations based on records. Only cost data were used, with no inclusion of benefits from each model, which were considered too diverse for robust quantification. Cost volume breakeven analysis is a common managerial tool to make comparisons between equipment or program alternatives [21]; using this approach, we compared the alternative models to identify the point at which they are similar in terms of simple costs.

Costs were separated according to fixed costs (the basic investment required to establish and implement data collection) and variable costs (those associated with each new group of women, covering all survey waves they will pass through). In reality, there are no separate groups of women for the census model, since recruitment is continuous; for the purposes of cost comparison, we identified the variable costs based on the number of women included in this study.

For the research team, the costs are incurred through survey development and testing, online survey building, user testing, implementation monitoring, design of the web platform, report building, and coordination. For health professionals, the costs are incurred through providing information to/inviting patients; enrolling patients, including data entry; monitoring results; and training in recruitment. For technology and infrastructure, costs are incurred through application programming interface connection development, maintenance of the survey platform, and maintaining the web platform to present data. Communication costs are incurred through provision of information to women and SMS invitations to women.

Statistical analysis was performed using Stata 15 and financial analysis was performed using Microsoft Excel.

Ethics Approval

The data collection was carried out within systematic surveys developed to monitor women's experiences, outcomes, and satisfaction with the Tuscan maternity pathways. As such, informed consent was not required, in line with the 2011 Italian guidelines on processing personal data to perform customer satisfaction surveys in the healthcare sector [22].

Results

Quantitative Results

A summary of the demographic characteristics for the two population groups is provided in Multimedia Appendix 1. Table 1 summarizes the main statistics for the survey responses.

As shown in Table 1, in the cohort model, 39% of women participated in the first survey wave, with 34% still participating in the final survey wave. Response rates for this model gradually reduced at each survey wave, reaching 87% after 1 year of enrollment. In the census model, 50% of women initially participated, falling to 23% in the final wave. Response rates reduced to 45% after 1 year in the model.

The participation rate represents the proportion of women completing survey waves out of the total eligible population. The eligible population for the cohort sample model is all women who gave birth in the relevant time period, according to hospital administrative data, irrespective of whether they joined the survey or not. For the census model, the eligible population is all women who received the pregnancy booklet and entered the maternal care pathway, irrespective of whether they joined the survey or not. The response rates in both models indicate the proportion of women responding to a survey who were successfully enrolled in the data collection model, completing the first survey (Table 1).

Table 1. Survey response descriptive statistics.

Statistic	Cohort model	Census model
Participation		
Total eligible women, N	9827	7826
Effective participation rate for first wave, n (%)	3849 (39.17)	3935 (50.28)
Effective participation rate for last survey wave, n (%)	3346 (34.05)	1788 (22.85)
Response rate, n (%)[a]		
T0/T0g	3849 (100.00)	3935 (100.00)
T1/T2g	3706 (96.28)	3038 (77.20)
T3/T3g	3633 (94.39)	2463 (62.59)
T5/T0p	3500 (90.93)	2325 (59.09)
T6/T1p	3477 (90.34)	1807 (45.92)
T12/T3p	3346 (86.93)	1788 (45.44)

[a]For the cohort model T0-T12 represent the time from delivery (0) and the months (1, 5, 6, and 12) postbirth. For the census model, T0g, T2g, and T3g represent the month (0, 2, and 3, respectively) of gestation, and T0p, T1p, and T3p represent the month (0, 1, and 3, respectively) postpartum. Also see Figure 1.

Regression Analysis of Survival Function

Univariate testing of variables indicated that all variables were relevant for further evaluation. No interaction terms were significant.

Testing the assumption of proportional hazards indicated that the impact of the data collection model was not proportional; the final regression model was thus stratified according to the data collection model. All other covariates followed the assumption of proportional hazards.
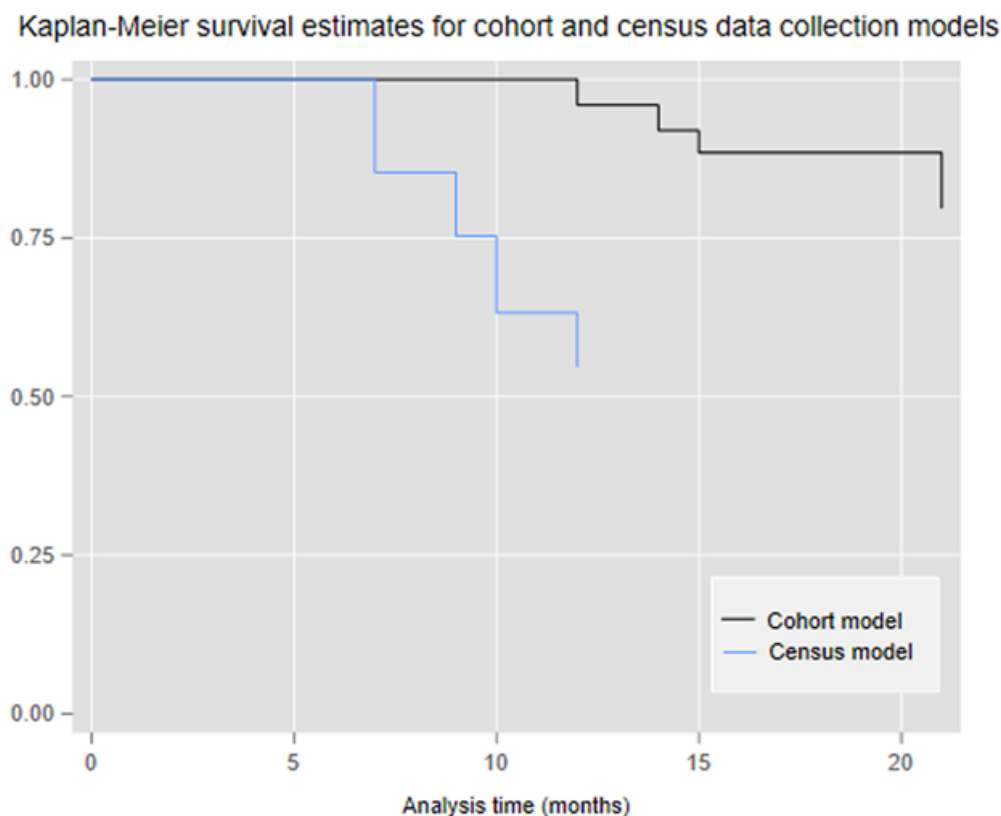
Stratification by Survey Type

As indicated in Table 2, all variables except having foreign citizenship were negatively associated with a failure event (not completing all survey waves). Increasing age showed a small reduction in the hazard ratio (HR) per year. Each education level had a reduction in hazard compared to the lowest level. Being employed compared with not employed and being in a relationship compared with being single were associated with a lower propensity to drop out. Women who previously had a child had a lower HR than women in their first pregnancy. It was not possible or appropriate to obtain a single value for the HR survey type on survival, since this changes over time.

As shown in the Kaplan-Meier plots for the two models (Figure 2), plots of the survival functions (not shown), and the descriptive statistics, the two data collection models showed different reductions in responses over time. The census collection model showed a large early drop in responses, followed by a less steep, but broadly steady, reduction over time, whereas the cohort model showed small, steady early reductions in responses, followed by a period of very limited reduction.

Table 2. Cox proportional hazards and multivariate hazard ratios.

Variable	Hazard ratio (SE)	95% CI	P value
Age	0.98 (0.00)	0.98-0.99	.001
Foreign	1.88 (0.11)	1.67-2.12	<.001
Education level			
High school diploma	0.77 (0.05)	0.68-0.88	<.001
Graduate	0.61 (0.04)	0.53-0.70	<.001
Employed	0.87 (0.05)	0.79-0.97	.01
In a relationship	0.84 (0.07)	0.70-0.99	.04
Nulliparous	0.86 (0.04)	0.79-0.95	.002

XSL•FO
RenderX

Figure 2. Kaplan-Meier curves of the two survey models.

Kaplan-Meier survival estimates for cohort and census data collection models

Qualitative Results

There were some common features of the models found, arising from the shared digital administration. These models can be used to collect and manage large volumes of patient-reported data. Such models also enable high response rates, as illustrated in the quantitative results. Additionally, both models are characterized by a comparatively high initial investment followed by low ongoing costs. Both models require analytical resources to derive insight from the data produced.

Digital administration also enables targeted surveys to be shared with expectant or new mothers according to their stage in the pathway, and need not be delivered alongside a specific intervention with a health care professional. This enables surveys to explore, in a timely manner, the aspects of experience or outcomes that are most relevant to people, rather than those based around institutions. This longitudinal design is uncommon in business-as-usual data collection. Additionally, under these models, both PROMs and PREMs can be collected separately or together, providing a more holistic picture of the dimensions of care than one data source in isolation.

There are also features of the two models that differ depending on the administration method, which are summarized in Table 3, categorized according to the relevance to methodology, management, and evaluation in the maternal pathway.

**Table 3.** Summary of methodological, managerial, and evaluative factors in each survey collection model.

Factors	Cohort model	Census model
Methodological factors		
Sample size	Medium to large sample size, predefined	Large, ever-growing sample size with ongoing recruitment
Representativeness of population	Based on deliveries in birth hospitals	Based on pregnancy at the district level, able to include women from small areas and those who give birth at home or in other settings
Survey timeliness	Survey at birth requires recall of experiences and outcomes during pregnancy	All surveys relating to the immediate preceding time period
Bias	Possible sampling bias: enrollment by health professionals after birth may encourage selection of mothers deemed to have had a more positive birth experience	Potential selection bias, although earlier recruitment of mothers reduces the risk of selection based around those deemed to have had a positive birth experience. Selection at first midwife appointment in pregnancy is blind to later experiences and outcomes
Collection burden	Need for staff training ahead of samples. Enrollment only needed up to a limited period, but is more time-consuming	Ongoing enrollment with less total time spent per health professional. Training only needed for new staff
Response rate	The initial effective response rate is high, (although lower than that of the census model), with low attrition	The initial effective response rate is the higher of the two models, although drops faster than that in the cohort model
PROMs[a] before/after birth	Pelvic floor PROMs are not included as there is no ability to collect prebirth data	Pelvic floor PROMs are included since baseline data at the beginning of the pregnancy are collected
Managerial factors		
Managerial insight	Data provide a snapshot of performance for a certain period of time, enabling lessons to be learned for the following period	Real-time data at different levels of geography enable targeted attention on areas where services need to work better or be better joined up
Health professional insights	Data provide a snapshot of performance for a certain period of time, enabling lessons to be learned for the following period	Possibility to provide real-time information to different care professionals about the state of delivery of care in their specific area, including highlighting where there are poor experiences or outcomes that professionals could address promptly through their activities
Evaluative factors		
Evaluation models	Enable multidimensional performance assessment	As in the cohort model, and additionally enable inclusion of patient-reported data alongside administrative measures, with contemporaneous reporting periods for both data sets
Evaluation periods	Data refer to a specific period of collection	Can be used "live" or at any given point in time for evaluating performance
Analytical approaches	Volume of data can be predetermined according to analytical requirements. Large data sets are possible, enabling advanced statistical models	Continuous collection enables additional analytical approaches (eg, difference in differences) to measure the impact of operational changes

[a]PROM: patient-reported outcome.

Cost Comparison

The fixed costs for the cohort and census surveys are calculated at €33,000 and €52,300 (US $1=€0.81 as of December 31, 2020), respectively. Variable costs per group of enrolled women are €18,040 and €15,360. There is therefore a higher initial cost and lower recurrent costs for the census survey compared with those of the cohort model. Projecting costs forward over several years showed that the point at which the census model becomes less costly overall is between years 7 and 8 (Figure 3).

XSL•FO

RenderX

Figure 3. Cost comparison of census and cohort models.
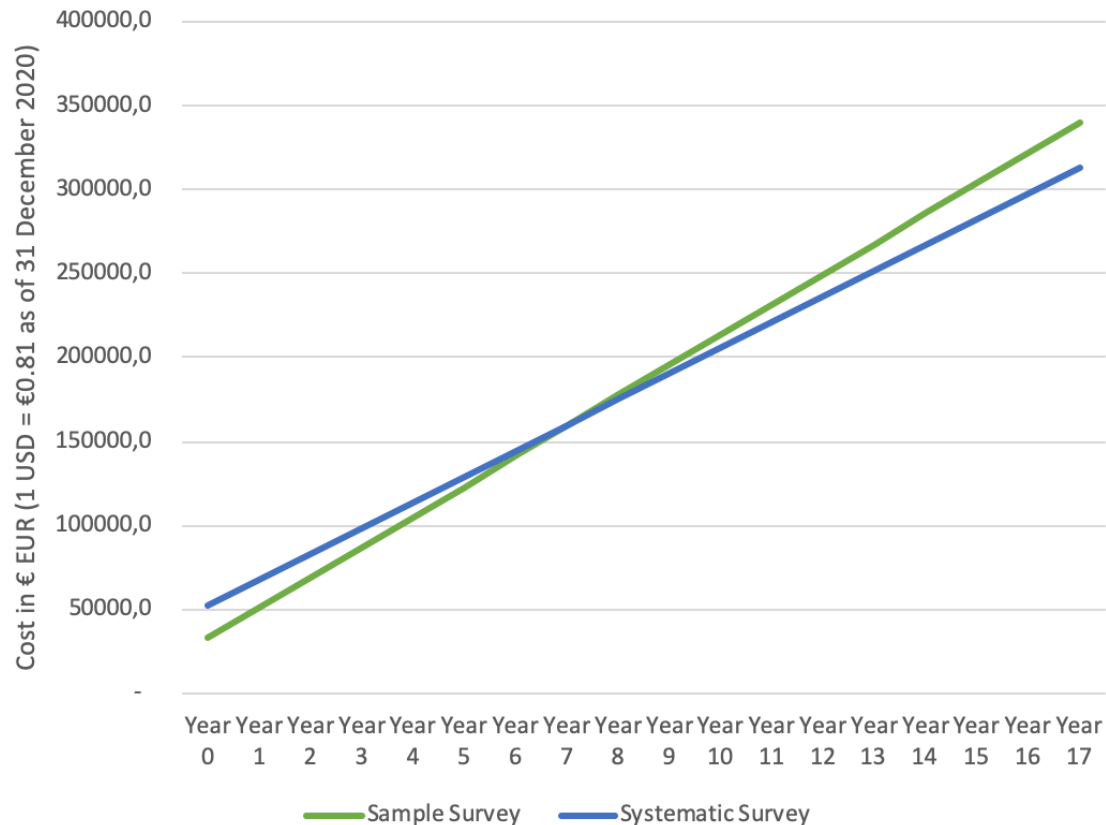
Discussion

Principal Results

We use mixed methods to evaluate the performance of two models of data collection for the maternal care pathway, both offering digital longitudinal collection of PROMs and PREMs. The models described, which are fully web-based with longitudinal collection of surveys targeted according to individuals' positions in the maternal pathway, are interesting from the perspectives of performance management, information, and implementation, and are also of international relevance.

The results of our analyses highlight that both models have some shared benefits. It is established that web-based surveys are cost-effective and provide the same measurement of variables as other collection methods, as well as additional benefits in completeness and data processing. A traditionally observed weakness in web-based surveys compared to postal surveys is their lower response rates (notwithstanding a broader trend of lower study participation across the board) [23-25]. This was not noted in the data collection models described in this study, which showed participation rates at the same level or higher than those of postal collection of maternal patient-reported data, and can be considered to be high in general terms for survey-based research, particularly for online surveys [17,24,26-28]. However, there remained a significant drop-off in responses over multiple survey waves. This could potentially be further improved by shortening the surveys or implementing other adjustments for user experience; further, more targeted feedback should be sought from participating mothers to identify the enablers and barriers to their continued participation in

multiple survey waves, particularly in the census model. Since the sample population of mothers is typically fairly young, this model of data collection likely avoids significant exclusion of respondents due to low digital capability as a result of age. However, as shown in the survival analysis, there remain systematic biases with respect to other population characteristics.

The findings from the multivariate regression showed that being less highly educated, not in a relationship, and unemployed were all associated with lower response rates, in line with results obtained over many years in research exploring the impact of participant characteristics on response likelihood [25,29,30]. We also found that having a foreign nationality was associated with a higher dropout risk. Although the data can be risk-adjusted to account for these factors when comparing different reporting areas or periods, it remains an unmet challenge to increase the representation of these groups in patient surveys. There is a risk that maternal services are providing poorer experiences and outcomes or are less responsive to the voices of disadvantaged women in particular, and that this is heightened by lower representation of such women in patient surveys.

The nonproportional hazard functions of the two survey models warrant further investigation to explore the extent to which the changing HRs over time are influenced by the surveys in question (ie, different surveys are differently acceptable and accessible for respondents), stage of pregnancy, and initial recruitment. One potential explanation is a greater selection bias of women in the cohort model (either through unconscious midwife selection or through self-selection, as previously observed in pregnancy cohorts [31]), leading to reduced dropout

rates in later stages. Our results suggest that the census collection model provides a more representative sample in the early survey waves, which reduces over time. This is supported by the higher initial participation rate in the census model (49.96%, 3935/7876), but with a lower final effective participation rate (22.70%, 1788/7876) 12 months after enrollment and response rate (45.44%, 1788/3935) after a total of six survey waves. The cohort model had a lower initial participation rate (39.17%, 3849/9827) and a remarkably low attrition rate, leading to an effective participation rate of 34.05% (3346/9827) after six survey waves and a response rate of 86.93% (3346/3849) 12 months after enrollment. The census model surveys are on average longer than the cohort model surveys, which may partially explain the higher attrition rates, although there was no apparent relationship between survey length and attrition within each model. These findings provide lessons for researchers and professionals seeking to collect the views of women in and around childbirth: the timing and mode of recruitment matter. This suggests that studies recruiting women exclusively around childbirth are subject to greater selection bias than those recruiting women earlier in pregnancy. A helpful development in reporting studies using data from new mothers would be to note the effective response rates (as in this study, with reference to the total population of women giving birth), rather than simply those who responded once invited.

The previously observed lower response rates in web surveys are typically based on models that include a postal element and are not digital-only models (ie, individuals are contacted by letter and then provided a web link to respond to the survey). Such mixed models necessarily require an additional step by survey respondents, rather than simply continuing to use the device on which they received the survey link. It is possible that some combination of the simpler, fully digital administration method and the previous in-person enrollment can lead to notably lower attrition rates (eg, 87% response rate at 12 months after enrollment in the cohort model) than have been achieved in other survey models.

From a health system performance intelligence perspective, the overall approach is noteworthy. The longitudinal PROM and PREM data collection in both models is new or uncommon in performance evaluation (typically such data may be collected for specific studies or for limited clinical use). Additionally, the delivery of different PREM and PROM surveys according to patients' stages in the pathway is a new development in performance measurement, unlike other longitudinal models of PROM collection where the same survey is given at multiple time points. In this way, the information collected is more relevant for assessment and improvement at each point. This could offer new options for using user-reported data in performance improvement, evaluations, and incentive models such as value-based purchasing or as an adjunct to bundled payments, to ensure the patient voice is given appropriate weight [32,33].

Strengths and Limitations

The application of survival analyses to survey waves is interesting and elucidating. The use of survival analysis to explore attrition in surveys was proposed by Eysenbach [34],

as fundamental to growing the "science of attrition," and has more recently been expounded upon in the context of web surveys [19]. Much of the published literature focuses on the methodology of attrition analyses, with particular attention paid to within-survey attrition [16]. Few studies have used survival methods to explore attrition across multiple survey waves [35] or have described the applied use of such methodologies to inform management practice and implementation. This study focused on the application of survival analysis to real-world survey data collected in multiple waves. This longitudinal experience and outcome data collection are pertinent and useful for measuring performance along a pathway, and can provide insights for managers and clinicians as well as researchers. This analysis thus informs both scholarship and practice in determining the most appropriate data collection models for different purposes.

Limitations of this study include that interpretation of statistical results is not straightforward, partly due to the nonproportional hazard functions of the two models so that a single HR for each model cannot be reported. Additionally, the nature of the multiple-wave survey models means that some women may have missed one or more survey waves without fully dropping out of the data collection. The impact of this is hard to capture. Some women may disengage from maternity pathway services after their initial encounter, resulting in an inaccurate population denominator or less accurate measurement of experiences. Some features of the surveys themselves can also affect response rates, which would require further investigation to distinguish from other model-dependent factors. Tuscan mothers were not included as lay representatives in development of the administration models (although they were involved through their roles as health care professionals and researchers); their involvement could provide further insights into the drivers of attrition (eg, on the impact of survey content).

The two populations of expectant and new mothers are also similar but not identical. These populations were recruited concurrently, not simultaneously, in different settings, and by different individuals (although by midwives in both cases). The demographic table for the two population groups in Multimedia Appendix 1 shows small but statistically significant differences, as would be expected for a large sample size as in this study. In particular, since the two data collection models commence at very different points in the maternal care pathway, women are exposed to different experiences and events at the same time postenrollment (ie, at 9 months), with one group giving birth while the other is caring for a 9-month-old child. This will likely result in different levels of willingness and ability to respond to surveys. Consequently, it is not advisable to make simple comparisons of response rates at the same time point postenrollment; response rates are likely determined by some interacting combination of time since enrollment, period in the maternal care pathway, and demographic characteristics. For example, it is notable that in the survey immediately postpartum, 4% of women in the cohort left the data collection model (1 month after enrollment), whereas 22% of women in the census left the collection (9 months after enrollment). In the first month of the census survey, 23% of women dropped out. There was no 9-month survey in the cohort model, but at 12 months only

a further 3% dropped out. These points support the observation that both the timing and mode of recruitment matter. The regression analysis controlled for population differences, but did not consider other factors. As such, in interpreting the results, it is necessary to consider the quantitative data (descriptive statistics, survival analysis, and regression) alongside the commentary about the stages in the pathway addressed by the time points and surveys in the two models.

Regarding the other methodological factors, some results are context-dependent. For example, costs are related to Italy and to the maternity pathway, and variable costs for the census model are imprecise. The qualitative benefits described are derived from the team's insights and relevant literature rather than from a full stakeholder review; thus, some may have been missed or inaccurately weighted.

Conclusions

Census collection has a wider set of potential benefits, particularly relating to use of the data as a management tool or for more granular performance evaluation. These benefits are shared by other clinical areas that broadly adopt this model of data collection [10]. However, the cost and effort are higher, which may not be justified in some use cases. A census model also has a higher dropout rate over time, necessitating increased methodological caution in later survey waves. The continuous collection model will be the more cost-effective approach in simple terms if the cohort sample data are collected more than seven times. Although the benefits were not quantified, it is clear that to fully realize the benefits possible from census data, professional and managerial action is required. This requires support including data translation, risk adjustment, and subsequent development of insight. Such efforts have costs. It is therefore probable that census data collection models are more appropriate only in health systems with fairly significant analytical capacity that are able to manage the data and support insight development *on an ongoing basis*. For occasional evaluations or for local areas starting to build their understanding of the reality of experience and outcomes for pregnant women and new mothers, a cohort sampling model may be more cost-effective. In the longer term, two emerging trends in health care will likely shift the balance toward the census model. First, the increasing sophistication of real-time automated analytics and decision-support tools for professionals will reduce the analytical resources required for continuously collected patient-reported data while simultaneously increasing their utility. Second, performance evaluation systems (and reimbursement models) are likely to give greater precedence to measures that matter the most to service users; in such situations, investment in systems akin to the census model will be a priority for all health systems.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Participant characteristics in the cohort and census sampling models.
[DOCX File , 19 KB - medinform_v10i3e25477_app1.docx]

References

1. Donabedian A. The quality of care. How can it be assessed? JAMA 1988;260(12):1743-1748. [doi: 10.1001/jama.260.12.1743] [Medline: 3045356]
2. Nuti S, Noto G, Vola F, Vainieri M. Let's play the patients music. Manag Decis 2018 Oct 08;56(10):2252-2272. [doi: 10.1108/md-09-2017-0907]
3. Porter ME. What is value in health care? N Engl J Med 2010 Dec 23;363(26):2477-2481. [doi: 10.1056/NEJMp1011024] [Medline: 21142528]
4. Kerr EA, Hayward RA. Patient-centered performance management: enhancing value for patients and health care systems. JAMA 2013 Jul 10;310(2):137-138 [FREE Full text] [doi: 10.1001/jama.2013.6828] [Medline: 23839743]
5. Defining value in "value-based healthcare." Report of the expert panel on effective ways of investing in health (EXPH). European Commission. 2019. URL: https://ec.europa.eu/health/system/files/2019-11/024_defining-value-vbhc_en_0.pdf [accessed 2020-09-12]

6.   Slawomirski L, Van Den Berg M. Harnessing the voice of the patient from the ward to the boardroom. World Hospitals and Health Services: The Official Journal of the International Hospital Federation. 2019. URL: https://www.uch.cat/documents/whhs-volume-54-number-3.pdf [accessed 2022-01-24]

7.   Black N, Burke L, Forrest CB, Sieberer UHR, Ahmed S, Valderas JM, et al. Patient-reported outcomes: pathways to better health, better services, and better societies. Qual Life Res 2016 May;25(5):1103-1112. [doi: 10.1007/s11136-015-1168-3] [Medline: 26563251]

8.   Coulter A, Fitzpatrick R, Cornwell J. The point of care. Measures of patients' experience in hospital: purpose, methods and uses. London: The King's Fund; 2009. URL: https://www.kingsfund.org.uk/sites/default/files/Point-of-Care-Measures-of-patients-experience-in-hospital-Kings-Fund-July-2009_0.pdf [accessed 2022-01-24]

9.   Bull C, Byrnes J, Hettiarachchi R, Downes M. A systematic review of the validity and reliability of patient-reported experience measures. Health Serv Res 2019 Oct;54(5):1023-1035 [FREE Full text] [doi: 10.1111/1475-6773.13187] [Medline: 31218671]

10.   De Rosis S, Cerasuolo D, Nuti S. Using patient-reported measures to drive change in healthcare: the experience of the digital, continuous and systematic PREMs observatory in Italy. BMC Health Serv Res 2020 Apr 16;20(1):315 [FREE Full text] [doi: 10.1186/s12913-020-05099-4] [Medline: 32299440]

11.   Escuriet R, White J, Beeckman K, Frith L, Leon-Larios F, Loytved C, EU COST Action IS0907. 'Childbirth Cultures, Concerns, Consequences'. Assessing the performance of maternity care in Europe: a critical exploration of tools and indicators. BMC Health Serv Res 2015 Nov 02;15:491 [FREE Full text] [doi: 10.1186/s12913-015-1151-2] [Medline: 26525577]

12.   Vedeler C, Nilsen A, Blix E, Downe S, Eri T. What women emphasise as important aspects of care in childbirth - an online survey. BJOG-Int J Obst Gy 2021 Sep 17:1-9. [doi: 10.1111/1471-0528.16926] [Medline: 34532959]

13.   Downe S, Finlayson K, Oladapo OT, Bonet M, Gülmezoglu AM. What matters to women during childbirth: A systematic qualitative review. PLoS One 2018;13(4):e0194906 [FREE Full text] [doi: 10.1371/journal.pone.0194906] [Medline: 29664907]

14.   WHO recommendations on antenatal care for a positive pregnancy experience. World Health Organization. 2016. URL: https://apps.who.int/iris/bitstream/handle/10665/250796/9789241549912-eng.pdf [accessed 2022-01-24]

15.   WHO recommendations: intrapartum care for a positive childbirth experience. World Health Organization. 2018. URL: https://www.who.int/reproductivehealth/publications/intrapartum-care-guidelines/en/ [accessed 2022-01-24]

16.   Blumenberg C, Zugna D, Popovic M, Pizzi C, Barros AJD, Richiardi L. Questionnaire breakoff and item nonresponse in web-based questionnaires: multilevel analysis of person-level and item design factors in a birth cohort. J Med Internet Res 2018 Dec 07;20(12):e11046 [FREE Full text] [doi: 10.2196/11046] [Medline: 30530454]

17.   Harrison S, Henderson J, Alderdice F, Quigley MA. Methods to increase response rates to a population-based maternity survey: a comparison of two pilot studies. BMC Med Res Methodol 2019 Mar 20;19(1):65 [FREE Full text] [doi: 10.1186/s12874-019-0702-3] [Medline: 30894130]

18.   Nuti S, De Rosis S, Bonciani M, Murante AM. Rethinking healthcare performance evaluation systems towards the people-centredness approach: their pathways, their experience, their evaluation. Healthc Pap 2017 Oct;17(2):56-64. [doi: 10.12927/hcpap.2017.25408] [Medline: 29595446]

19.   Hochheimer CJ, Sabo RT, Krist AH, Day T, Cyrus J, Woolf SH. Methods for evaluating respondent attrition in web-based surveys. J Med Internet Res 2016 Nov 22;18(11):e301 [FREE Full text] [doi: 10.2196/jmir.6342] [Medline: 27876687]

20.   de Haan-Rietdijk S, Voelkle MC, Keijsers L, Hamaker EL. Discrete- vs. continuous-time modeling of unequally spaced experience sampling method data. Front Psychol 2017;8:1849. [doi: 10.3389/fpsyg.2017.01849] [Medline: 29104554]

21.   Finkler S, Ward D, Baker J. Essentials of Cost Accounting for Health Care Organizations. Burlington: Jones & Bartlett Learning; 1999.

22.   Guidelines on processing personal data to perform customer satisfaction surveys in the health care sector. The Italian Data Protection Authority. 2011. URL: https://www.garanteprivacy.it/home/docweb/-/docweb-display/docweb/3853781 [accessed 2020-06-15]

23.   van Gelder MMHJ, Bretveld RW, Roeleveld N. Web-based questionnaires: the future in epidemiology? Am J Epidemiol 2010 Dec 01;172(11):1292-1298. [doi: 10.1093/aje/kwq291] [Medline: 20880962]

24.   Whitehead L. Methodological issues in internet-mediated research: a randomized comparison of internet versus mailed questionnaires. J Med Internet Res 2011 Dec 04;13(4):e109 [FREE Full text] [doi: 10.2196/jmir.1593] [Medline: 22155721]

25.   Galea S, Tracy M. Participation rates in epidemiologic studies. Ann Epidemiol 2007 Sep;17(9):643-653. [doi: 10.1016/j.annepidem.2007.03.013] [Medline: 17553702]

26.   Redshaw M, Heikkila K. Delivered with care: a national survey of women's experience of maternity care 2010. National Perinatal Epidemiology Unit, University of Oxford. URL: https://www.npeu.ox.ac.uk/assets/downloads/reports/Maternity-Survey-Report-2010.pdf [accessed 2020-10-10]

27.   Redshaw M, Henderson J. Safely delivered: a national survey of women's experience of maternity care. National Perinatal Epidemiology Unit. 2014. URL: https://www.npeu.ox.ac.uk/assets/downloads/reports/Safely%20delivered%20NMS%202014.pdf [accessed 2020-12-10]

28.   Guo Y, Kopec JA, Cibere J, Li LC, Goldsmith CH. Population survey features and response rates: a randomized experiment. Am J Public Health 2016 Aug;106(8):1422-1426. [doi: 10.2105/AJPH.2016.303198] [Medline: 27196650]

29.   Gustavson K, von Soest T, Karevold E, Røysamb E. Attrition and generalizability in longitudinal studies: findings from a 15-year population-based study and a Monte Carlo simulation study. BMC Public Health 2012 Oct 29;12:918 [FREE Full text] [doi: 10.1186/1471-2458-12-918] [Medline: 23107281]

30.   Jang M, Vorderstrasse A. Socioeconomic status and racial or ethnic differences in participation: web-based survey. JMIR Res Protoc 2019 Apr 10;8(4):e11865 [FREE Full text] [doi: 10.2196/11865] [Medline: 30969173]

31.   Nilsen RM, Vollset SE, Gjessing HK, Skjaerven R, Melve KK, Schreuder P, et al. Self-selection and bias in a large prospective pregnancy cohort in Norway. Paediatr Perinat Epidemiol 2009 Nov;23(6):597-608. [doi: 10.1111/j.1365-3016.2009.01062.x] [Medline: 19840297]

32.   Trombley MJ, McClellan SR, Kahvecioglu DC, Gu Q, Hassol A, Creel AH, et al. Association of Medicare's bundled payments for care improvement initiative with patient-reported outcomes. Health Serv Res 2019 Aug;54(4):793-804. [doi: 10.1111/1475-6773.13159] [Medline: 31038207]

33.   Schlesinger M, Grob R, Shaller D. Using patient-reported information to improve clinical practice. Health Serv Res 2015 Dec;50(Suppl 2):2116-2154 [FREE Full text] [doi: 10.1111/1475-6773.12420] [Medline: 26573890]

34.   Eysenbach G. The law of attrition. J Med Internet Res 2005 Mar 31;7(1):e11 [FREE Full text] [doi: 10.2196/jmir.7.1.e11] [Medline: 15829473]

35.   Zethof D, Nagelhout GE, de Rooij M, Driezen P, Fong GT, van den Putte B, et al. Attrition analysed in five waves of a longitudinal yearly survey of smokers: findings from the ITC Netherlands survey. Eur J Public Health 2016 Aug;26(4):693-699 [FREE Full text] [doi: 10.1093/eurpub/ckw037] [Medline: 27060589]

Abbreviations

GLMM: generalized linear mixed model
HR: hazard ratio
PREM: patient-reported experience measure
PROM: patient-reported outcome measure

Original Paper

Using a Convolutional Neural Network and Convolutional Long Short-term Memory to Automatically Detect Aneurysms on 2D Digital Subtraction Angiography Images: Framework Development and Validation

JunHua Liao[1,2*], BSc; LunXin Liu[1*], MD; HaiHan Duan[3], MSc; YunZhi Huang[4], PhD; LiangXue Zhou[1], MD; LiangYin Chen[2], PhD; ChaoHua Wang[1], MD

[1]Department of Neurosurgery, West China Hospital, Sichuan University, Chengdu, China

[2]College of Computer Science, Sichuan University, Chengdu, China

[3]School of Science and Engineering, The Chinese University of Hong Kong, Shenzhen, China

[4]School of Automation, Nanjing University of Information Science and Technology, Nanjing, China

*these authors contributed equally

Corresponding Author:
ChaoHua Wang, MD
Department of Neurosurgery
West China Hospital
Sichuan University
No. 37 Guoxue Lane, Wuhou District
Chengdu, 610041
China
Phone: 86 18628169123
Email: wangchaohuaHX@163.com

Abstract

Background: It is hard to distinguish cerebral aneurysms from overlapping vessels in 2D digital subtraction angiography (DSA) images due to these images' lack of spatial information.

Objective: The aims of this study were to (1) construct a deep learning diagnostic system to improve the ability to detect posterior communicating artery aneurysms on 2D DSA images and (2) validate the efficiency of the deep learning diagnostic system in 2D DSA aneurysm detection.

Methods: We proposed a 2-stage detection system. First, we established the region localization stage to automatically locate specific detection regions of raw 2D DSA sequences. Second, in the intracranial aneurysm detection stage, we constructed a bi-input+RetinaNet+convolutional long short-term memory (C-LSTM) framework to compare its performance for aneurysm detection with that of 3 existing frameworks. Each of the frameworks had a 5-fold cross-validation scheme. The receiver operating characteristic curve, the area under the curve (AUC) value, mean average precision, sensitivity, specificity, and accuracy were used to assess the abilities of different frameworks.

Results: A total of 255 patients with posterior communicating artery aneurysms and 20 patients without aneurysms were included in this study. The best AUC values of the RetinaNet, RetinaNet+C-LSTM, bi-input+RetinaNet, and bi-input+RetinaNet+C-LSTM frameworks were 0.95, 0.96, 0.92, and 0.97, respectively. The mean sensitivities of the RetinaNet, RetinaNet+C-LSTM, bi-input+RetinaNet, and bi-input+RetinaNet+C-LSTM frameworks and human experts were 89% (range 67.02%-98.43%), 88% (range 65.76%-98.06%), 87% (range 64.53%-97.66%), 89% (range 67.02%-98.43%), and 90% (range 68.30%-98.77%), respectively. The mean specificities of the RetinaNet, RetinaNet+C-LSTM, bi-input+RetinaNet, and bi-input+RetinaNet+C-LSTM frameworks and human experts were 80% (range 56.34%-94.27%), 89% (range 67.02%-98.43%), 86% (range 63.31%-97.24%), 93% (range 72.30%-99.56%), and 90% (range 68.30%-98.77%), respectively. The mean accuracies of the RetinaNet, RetinaNet+C-LSTM, bi-input+RetinaNet, and bi-input+RetinaNet+C-LSTM frameworks and human experts were 84.50% (range 69.57%-93.97%), 88.50% (range 74.44%-96.39%), 86.50% (range 71.97%-95.22%), 91% (range 77.63%-97.72%), and 90% (range 76.34%-97.21%), respectively.

XSL•FO
RenderX

Conclusions: According to our results, more spatial and temporal information can help improve the performance of the frameworks. Therefore, the bi-input+RetinaNet+C-LSTM framework had the best performance when compared to that of the other frameworks. Our study demonstrates that our system can assist physicians in detecting intracranial aneurysms on 2D DSA images.

Introduction

The prevalence of cerebral aneurysms in the general population is approximately 2% to 3% [1]. When an intracranial aneurysm ruptures, it may bleed into the brain parenchyma, causing a hemorrhage of the cerebral parenchyma, or, more commonly, it bleeds into the subarachnoid space and causes a subarachnoid hemorrhage (SAH). An SAH is a catastrophic event with a mortality rate of 25% to 50%. Nearly 50% of SAH survivors have permanent disabilities; only approximately one-third of patients with SAH have good prognoses [2,3]. Hence, it is crucial to detect and treat aneurysms as early as possible. The gold standard for diagnosing cerebral aneurysms is digital subtraction angiography (DSA). The application of 3D DSA has dramatically improved the diagnostic accuracy for aneurysms. However, as many hospitals lack the technical and reconstitution equipment for 3D DSA, especially in low-income countries, radiologists in these hospitals have to diagnose cerebral aneurysms by using 2D DSA images. Unlike 3D images, 2D DSA images lack spatial information, and it is difficult for radiologists to distinguish aneurysms from overlapping vessels in 2D DSA images. Therefore, the assessment of these 2D DSA images is usually subjective and may be influenced by the experience of radiologists.

In recent years, image recognition via deep learning for diagnostic imaging has achieved good performance in various medical fields, such as skin cancer, retinopathy, pneumonia, and gastric cancer [4]. Deep learning represents a new machine learning method that enables a machine to analyze various training images, so that it can extract specific clinical features [5]. Based on the cumulative clinical features, a machine can prediagnose newly acquired clinical images.

A convolutional neural network (CNN) is a type of deep learning model for processing data that have a grid pattern, such as images. CNNs were inspired by the organization of the animal visual cortex [6,7] and designed to automatically and adaptively learn spatial hierarchies of characteristics from low- to high-level pictures. CNNs have achieved good performance in several medical fields, such as lesion detection [8] and classification [9].

Convolutional long short-term memory (C-LSTM) networks [10] have advantages over feedforward neural networks, as they can discover the hidden structures of medical time signals.

C-LSTM networks can perform pattern recognition analyses on medical time series data and have obtained high accuracies in the classification of medical signals [11,12].

Recent studies have used deep learning methods for detecting cerebral aneurysms in 2D DSA images, but these works have some limitations. Podgoršak et al [13] modified the Visual Geometry Group network—a network used for classification—into a network suitable for semantic segmentation tasks for detecting aneurysms. The data set of their study was composed of positive case data for aneurysms, and its false-positive rate has not been evaluated. Jin et al [14] used a bidirectional C-LSTM network to segment aneurysms; although the network's patient-level sensitivity was 97.7%, the average number of false positives per sequence was as high as 3.77. Liao et al [15] used a C-LSTM network to extract time information when detecting aneurysms but did not consider the relationships among DSA images from different aspects of the same patient. Duan et al [16] combined frontal and lateral DSA images for detection but did not use the timing information of the DSA sequence. This method requires an additional false-positive correction algorithm for correcting the results. Therefore, the existing deep learning–based aneurysm detection methods still need to be improved.

To solve the aforementioned problems, we combined a CNN for acquiring spatial information and a C-LSTM network for learning temporal information to detect aneurysms in 2D DSA images.

Posterior communicating artery (PCoA) aneurysms are one of the most common aneurysms encountered by neurosurgeons and neurointerventional radiologists and are the second most common aneurysms overall (25% of all aneurysms), representing 50% of all internal carotid artery (ICA) aneurysms [17]. Hence, to solve the problem of data deficiency, we focused on PCoA aneurysms to (1) construct a deep learning diagnostic system to improve the ability to detect PCoA aneurysms on 2D DSA images and (2) validate the efficiency of the deep learning diagnostic system in 2D DSA aneurysm detection.

This deep learning diagnostic system includes a region localization stage (RLS) and an intracranial aneurysm detection stage (IADS). The RLS is used to automatically locate a specific detection area, and in the IADS, the system conducts aneurysm detection for the area images outputted in the RLS. The cascading framework flowchart is shown in Figure 1.

Figure 1. The flowchart of the deep learning diagnostic system. "X^i_t" represents the tth frame in the digital subtraction angiography sequence of the ith patient. C-LSTM: convolutional long short-term memory; FPN: feature pyramid network; IADS: intracranial aneurysm detection stage; ResNet: residual deep neural network; RLS: region localization stage.
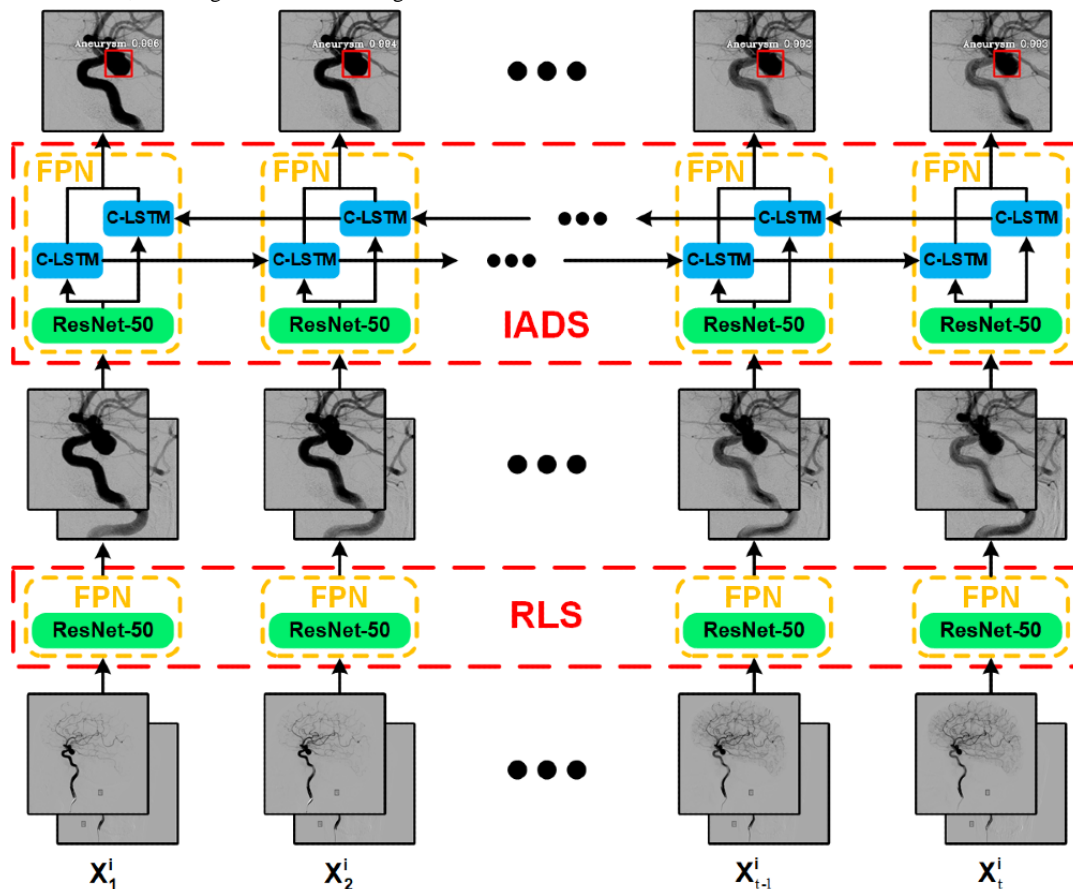
The main contributions of this paper can be summarized as follows. First, the bi-input network framework of the IADS increases the amount of information and then combines spatial-temporal information through feature pyramid networks (FPNs) [18], with a residual deep neural network (ResNet) [19] and bidirectional C-LSTM network acting as the backbone. This greatly improves the accuracy and efficiency of aneurysm detection. Second, our proposed method can achieve low false-positive rates without the need for a false-positive correction algorithm.

Methods

Ethics Approval

This retrospective study was approved (number 20220310005) by the institutional review board of West China Hospital, Sichuan University, Chengdu, China, with a waiver of written informed consent.

Study Design

A total of 586 patients who underwent DSA examination and had identified PCoA aneurysms from January 2014 to December 2019 in West China Hospital were included in this study. All of the PCoA aneurysms were double confirmed via 3D DSA. The main inclusion criterion stipulated that patients were diagnosed with PCoA aneurysms via DSA. The exclusion criteria consisted of the following: (1) patients lacking lateral

frontal DSA images; (2) patients with arteriovenous malformations, arteriovenous fistulas, or moyamoya disease; (3) patients with treated aneurysms; and (4) patients with aneurysms in other locations.

The obtained images were in DICOM format, which requires a large memory space. To decrease the computational load and improve usability, we converted the images to PNG format in model training and testing.

Two experienced radiologists identified 6 to 12 frameworks for 2D DSA images, which provided sufficient visualization of the PCoA region. Manual annotations were performed for the identification of aneurysms, vessel overlaps, and PCoA regions. To augment the training data, each image was rotated randomly between 0° and 359°. The data set was divided into the following three parts: the training set, validation set, and test set. The training set was used to train the algorithm, the validation set was used for model selection, and the test set was used for the assessment of the final chosen model. To obtain a reliable and stable model, this study adopted 5-fold cross-validation, during which the data set was divided into 5 parts; 4 parts were used for training and 1 part was used for validation. The mean value of the 5 results was used as the algorithm accuracy. The advantage of cross-validation is that it can make full use of limited data to find suitable model parameters and prevent overfitting. Raw 2D DSA images usually have large resolutions. Initially, the detection of intracranial aneurysms was based on original 2D DSA images, and the large resolution of the original

2D DSA images resulted in extra time consumption and interference. Specifically, researchers have attempted to avoid large resolution–related problems by manually locating detection areas requiring considerable amounts of work. In our case, we used the RLS to automatically locate specific detection regions of raw 2D DSA sequences, as shown in Figure 2. This method can be used to reduce the interference in aneurysm detection. In theory, region localization can be performed to locate any ICA region, but we could only prove the feasibility of using the RLS to identify PCoA regions due to the limitations of the data set. As shown in Figure 2, this architecture uses a raw 2D DSA

sequence as input. The ResNet-50–based [19] FPN sends the features extracted from each frame to anchor boxes [20] to predict the PCoA region. The detector outputs 6n parameters in which "6" represents the bounding box's x-coordinate, y-coordinate, width, height, classification label, and confidence for classification and "n" refers to the n objects detected in the RLS. The bounding box with the highest prediction confidence was applied to other frameworks in the DSA sequence, and it outputted the PCoA region sequence. Moreover, to connect with the IADS, each frame of the output sequence was resized to 288×288 pixels during the RLS.

Figure 2. The network architecture of the RLS. "Conv f×f, c, /s" represents a 2D convolutional layer with a kernel size of f×f, a c number of channels, and an s number of strides, which is defaulted to 1. "Pool f×f, /s" denotes the maximum pooling layer, which has a filter size of f and an s number of strides. The "anchor" is used to predict the PCoA region, and "up-sampling" refers to nearest neighbor up-sampling with an up-sampling rate of 2. PCoA: posterior communicating artery; RLS: region localization stage.
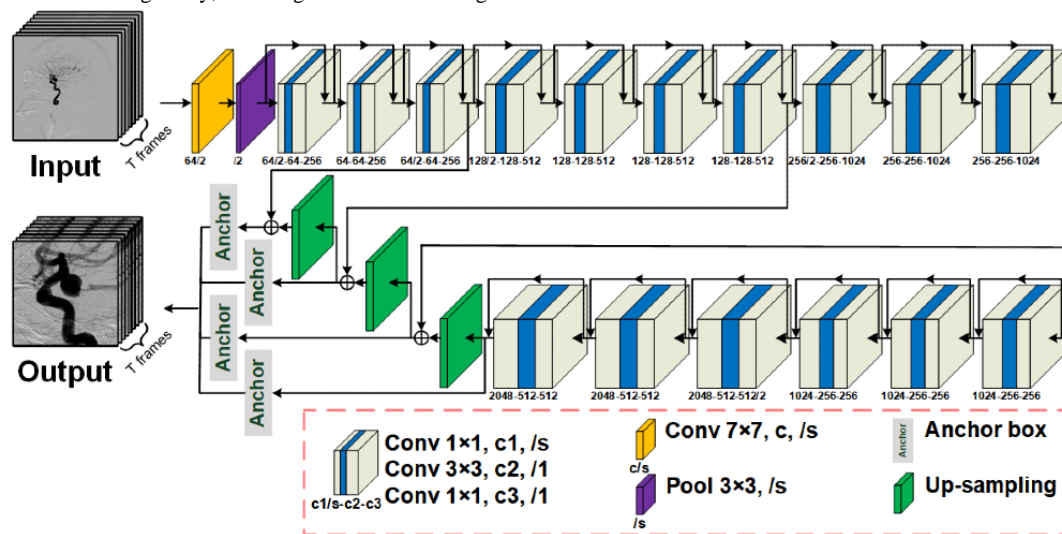
ResNet was the winner of the 2015 ImageNet Large Scale Visual Recognition Challenge for image classification [19]. It has several advantages over traditional CNNs, as follows: (1) it accelerates the training speed of deep networks; (2) instead of widening the network, it increases the depth of the network, resulting in fewer extra parameters; (3) the residual block inside ResNet uses jump connections to alleviate the problem of gradient disappearance resulting from the increase in the depth of the deep neural network; and (4) it achieves higher accuracy in network performance, especially in image classification [19]. Due to the excellent performance of ResNet, it has been widely used in various medical imaging tasks [21-23].

C-LSTM is a variant of long short-term memory (LSTM) that has a convolution operation inside of the LSTM cell. Both models are special kinds of recurrent neural networks that are capable of learning long-term dependencies. The main difference between C-LSTM and LSTM is the number of input dimensions. Using LSTM to process image sequences with temporal information requires converting 3D data to 2D data, which inevitably results in the loss of information. C-LSTM networks inherit the advantages of traditional LSTM networks and are very suitable for the analysis of spatiotemporal data due to their internal convolution structure. Therefore, many studies use C-LSTM to process medical image sequences [11,12,24].

Detecting objects at different scales, particularly small objects, is challenging. FPNs combine low-resolution, semantically strong features with high-resolution, semantically weak features via a top-down pathway and lateral connections. FPNs have rich semantics at all levels and are built quickly from a single-input image scale without sacrificing representational power, speed, or memory [18].

Object detection algorithms have 2 classic structures—1-stage and 2-stage algorithms. Compared to the 1-stage algorithm, the 2-stage algorithm has 1 more step for solving the problem of class imbalance. Therefore, the 2-stage algorithm is more time-consuming. Lin et al [25] constructed RetinaNet by combining ResNet, FPNs, and fully convolutional networks [26]. The RetinaNet algorithm solves the problem of class imbalance by using the focal loss function instead of the proposal extraction step, thereby greatly improving the detection speed with high accuracy. Because of the excellent performance of RetinaNet, it is widely used in object detection tasks involving medical images [27-29].

We compared the following three structures in the IADS: (1) RetinaNet [25], which uses single-frame images as input; (2) RetinaNet+C-LSTM [15], which is based on RetinaNet and uses C-LSTM to extract bidirectional time information and take frontal or lateral DSA sequences as input; and (3)

bi-input+RetinaNet [16], which combines frontal and lateral DSA sequences together as input.

As shown in Figure 3, the target sequence of the PCoA region and its corresponding frontal or lateral sequence were concatenated as a 6-channel image sequence in which the target sequence occupies the first 3 channels. ResNet-50 extracted individual spatial features from each 6-channel frame in the input sequence. In total, 3 feature layers were selected for temporal feature extraction by using bidirectional C-LSTM, namely C3, C4, and C5, which had 512, 1024, and 2048

channels, respectively. It should be noted that the number of channels in the C-LSTM network was set as half of the input. After extracting the temporal information, we concatenated the features of the forward C-LSTM network and the reverse C-LSTM network and sent them to the FPN for further extraction. Anchor boxes identified intracranial aneurysms and overlapping blood vessels based on the features extracted by the FPN. To make the detection results more reliable, the detector only outputted the predicted objects with a confidence level of >0.6.

Figure 3. The network architecture of the IADS. "X_t^i" represents the tth frame in the DSA sequence of the ith patient. "Conv: f×f, /s" represents a convolutional layer with a kernel size of f×f and an s number of strides, where s is defaulted to 1. The channel of the convolutional layer defaults to 256. "C3," "C4," and "C5" represent the 3-layer features of ResNet-50. "Up-sample" refers to nearest neighbor up-sampling with an up-sampling rate of 2. The "anchor" denotes the anchor box, which uses the features to output the detection result. C-LSTM: convolutional long short-term memory; Conv 2D: 2D convolution; DSA: digital subtraction angiography; FPN: feature pyramid network; IADS: intracranial aneurysm detection stage; ResNet: residual deep neural network.
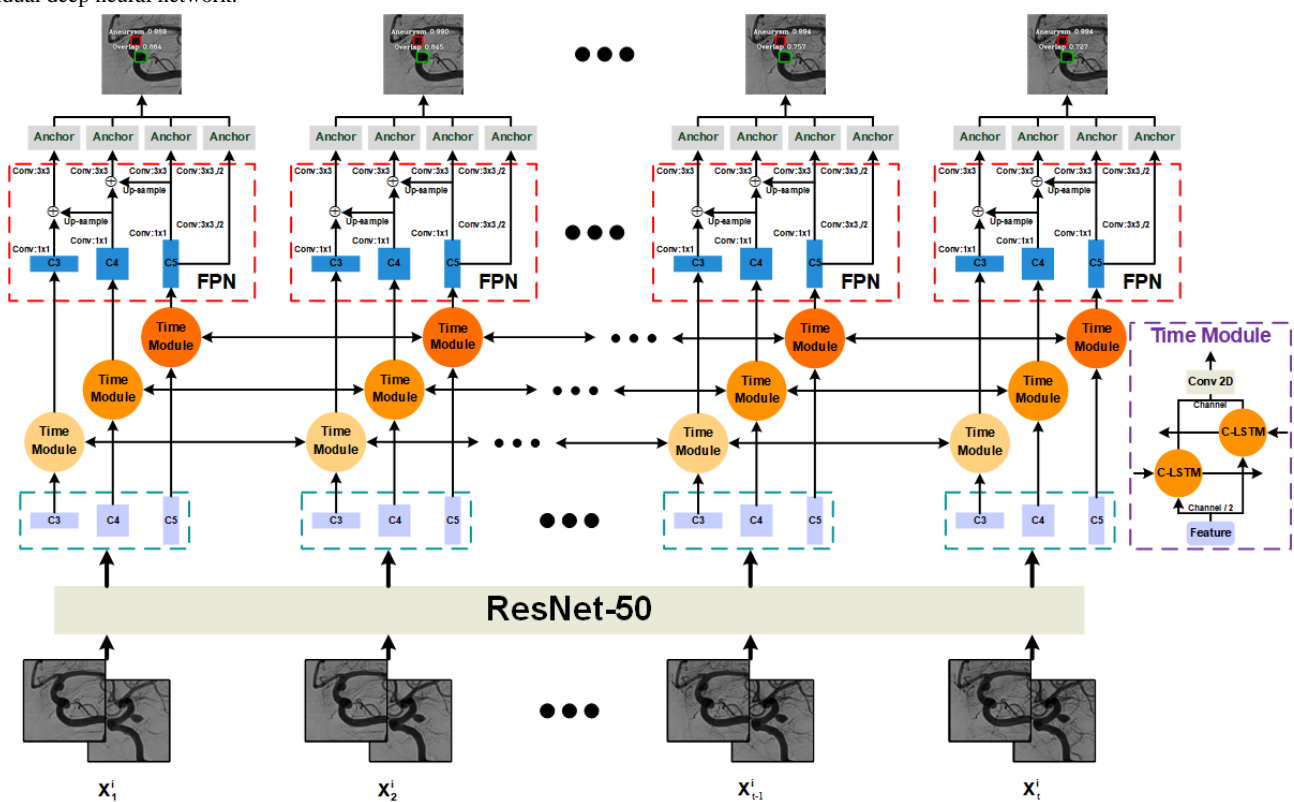
All models were trained and tested with a Keras [30] deep learning framework on an NVIDIA GTX 1080Ti graphics processing unit (11GB GDDR5X; NVIDIA Corporation). We used the data in the training set to train the region localization and intracranial aneurysm detection algorithms, and the initial learning rate of each step in the training process was set to 3×10^{-6} for the RLS and 1×10^{-4} for the IADS. The Adam optimization method [31] was adopted, and the learning rate was dynamically adjusted with the training progress. If the variation in the range of loss in 2 consecutive epochs was less than 1×10^{-4}, then the learning rate was reduced by a factor of 10. This method achieved the local optimum of the training process.

The loss function for object classification used focal loss [25]. This loss function reduced the weight of the large number of simple negative samples in training, thereby solving the problem

of a serious imbalance in the ratio of positive to negative samples in object detection tasks. The focal loss was defined as follows:

where "FL" denotes focal loss, "α" denotes the balanced parameter used to balance the proportional inequality of positive and negative samples, "γ" denotes the downweighted rate, "p" represents prediction confidence, and "y {±1}" is the ground truth class. When γ was >0, the loss function reduced the loss of easy-to-classify samples and thus focused more on difficult and misclassified samples. Specifically, we used an α of .25 and a γ of 2.0 in the training process.

Smooth L1 loss [25] was used as the loss function for bounding box regression. As a commonly used loss function in regression tasks, smooth L1 loss can limit the gradient value from the

following two aspects to prevent training failure: (1) when the difference between the predicted value and the ground truth was too large, the gradient value was not too large, and (2) when the predicted value was very close to the ground truth, the gradient value was small enough. This loss function was defined as follows:

in which

where "SL" denotes smooth L1 loss, "t" denotes the bounding box of the predicted object, "v" represents the bounding box of the ground truth, and "σ" is the weighted factor. A σ of 3.0 was used in the training process.

Statistical Analysis

Statistical analyses were performed by using statistical software (SPSS version 22.0; IBM Corporation). We used the 5-fold cross-validation strategy with mean average precision (mAP) values to assess the accuracy of intracranial aneurysm and overlap classification. The bounding box regression task was evaluated based on the smooth L1 loss. A confusion matrix, receiver operating characteristic (ROC) curves, and area under the curve (AUC) values were used to assess the abilities of different frameworks. For ROC curves, comparisons of AUC values (with SEs and 95% CIs) were made by using a nonparametric approach [32]. A total of 20 patients with PCoA aneurysms (test set) and 20 patients without aneurysms were used to evaluate the performance of each framework and the human experts, who had 20 years of experience. True positives, true negatives, false positives, and false negatives were used to calculate sensitivity, specificity, and accuracy, which were determined based on the optimal threshold from the Youden index. The adjusted Wald method was used to determine the 95% CIs of the accuracy, sensitivity, and specificity values from the contingency tables [33].

Results

During the RLS, the system only needs to perform the simple task of determining the valid coarse regions. The accuracy of region localization for the test set was 100%, which proves that

this method accurately located the PCoA regions from the original DSA images.

Of the 275 patients included in this study, 255 had PCoA aneurysms, and 20 did not have aneurysms. A flowchart of the enrolled patients is shown in Figure 4.

The AUC values and the ROC curves of RetinaNet [25], Liao et al [15], Duan et al [16], and the bi-input+RetinaNet+C-LSTM framework are shown in Figure 5. The focal loss and the smooth L1 loss also showed that the aforementioned frameworks had sufficient convergence (Figures 6 and 7). Compared to the average AUC values of RetinaNet [25] (0.920), Liao et al [15] (0.920) and Duan et al [16] (0.916), the bi-input+RetinaNet+C-LSTM framework had the largest average AUC value (0.936). The 5-fold cross-validation mAP values of the aforementioned frameworks are listed in Table 1. The mAP represents the average area under the precision-recall curves that were drawn based on the results of aneurysm and blood vessel overlap predictions.

The sensitivity, specificity, and accuracy results of RetinaNet [25], Liao et al [15], Duan et al [16], the bi-input+RetinaNet+C-LSTM framework, and the human experts with 20 years of experience are listed in Table 2.

Compared to the other frameworks' results, the bi-input+RetinaNet+C-LSTM framework had the best performance. The mean sensitivity, specificity, and accuracy of the bi-input+RetinaNet+C-LSTM framework were 89% (range 67.02%-98.43%), 93% (range 72.30%-99.56%), and 91% (range 77.63%-97.72%), respectively.

The confusion matrix of each framework is shown in Figure 8; both the bi-input+RetinaNet+C-LSTM and RetinaNet frameworks had the highest true-positive rate, but the false-positive rate of the bi-input+RetinaNet+C-LSTM framework was much smaller than that of the other frameworks. Therefore, the bi-input+RetinaNet+C-LSTM framework had the best performance compared to that of the other frameworks.

The original images of the DSA sequence and their corresponding results for the RLS and IADS are presented in Figure 9, which shows the detection results for different sizes of aneurysms. Most of the results had a confidence level of up to 1.0. This proves that our proposed method performs well in the detection of multiscale aneurysms.

Figure 4. Flowchart of enrollment information for included patients. AVF: arteriovenous fistula; AVM: arteriovenous malformation; DSA: digital subtraction angiography; PCoA: posterior communicating artery.
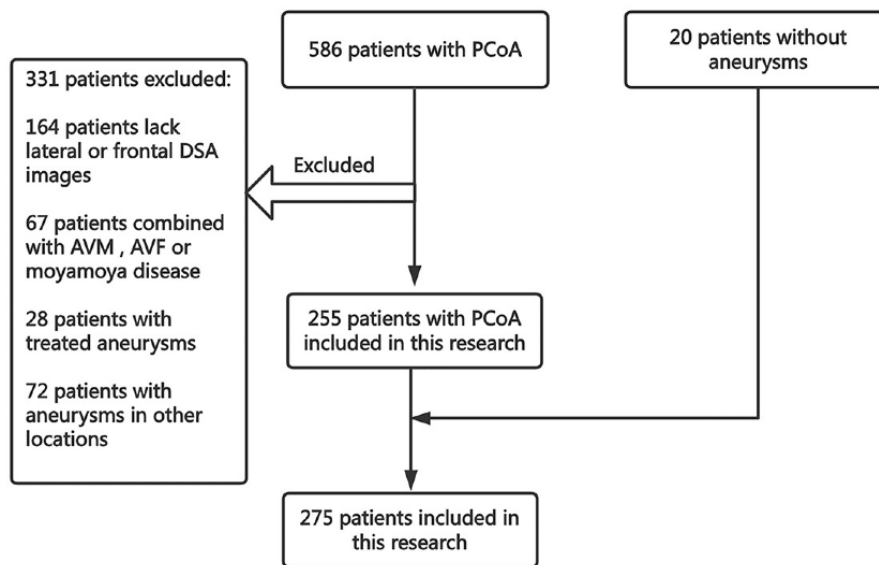
Figure 5. The 5-fold cross-validation results for the ROC curves and AUC values of the different frameworks. The results of different cross-validation models are shown in different colors. A: RetinaNet [25]. B: Liao et al [15]. C: Duan et al [16]. D: Bi-input+RetinaNet+C-LSTM. The ROC curves of fold 0 and fold 2 in graph C overlap, and the ROC curves of fold 1 and fold 4 in graph D overlap. AUC: area under the curve; C-LSTM: convolutional long short-term memory; ROC: receiver operating characteristic.
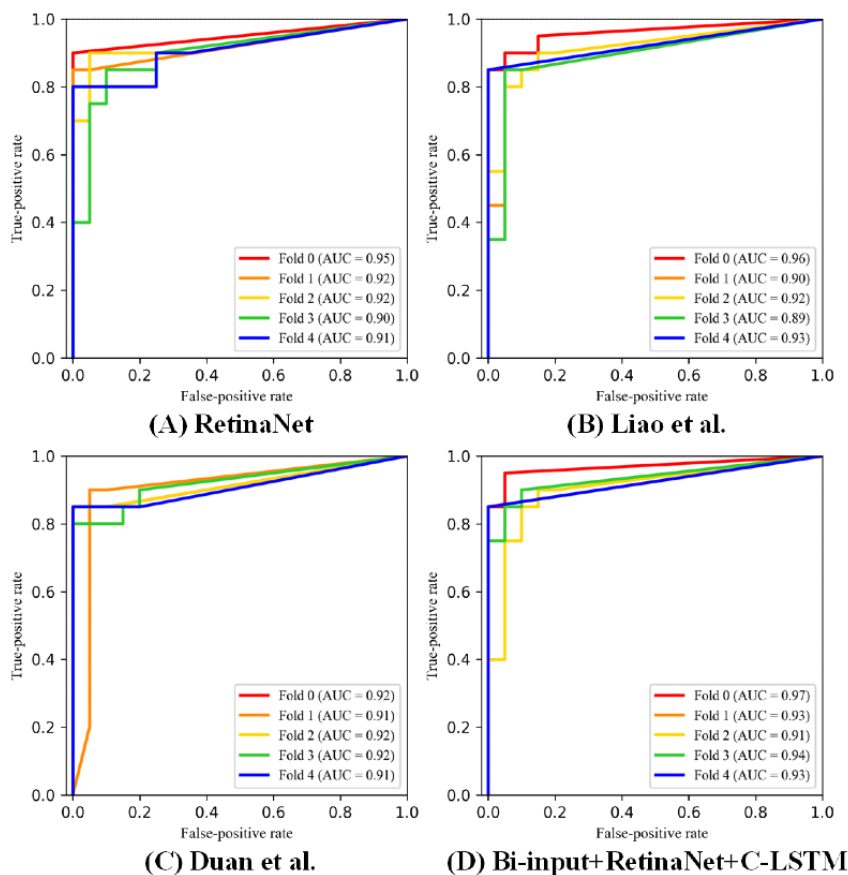
Figure 6. The 5-fold cross-validation results of the focal loss of each framework. Different color curves indicate different cross-validation models. A: RetinaNet [25]. B: Liao et al [15]. C: Duan et al [16]. D: Bi-input+RetinaNet+C-LSTM. C-LSTM: convolutional long short-term memory.
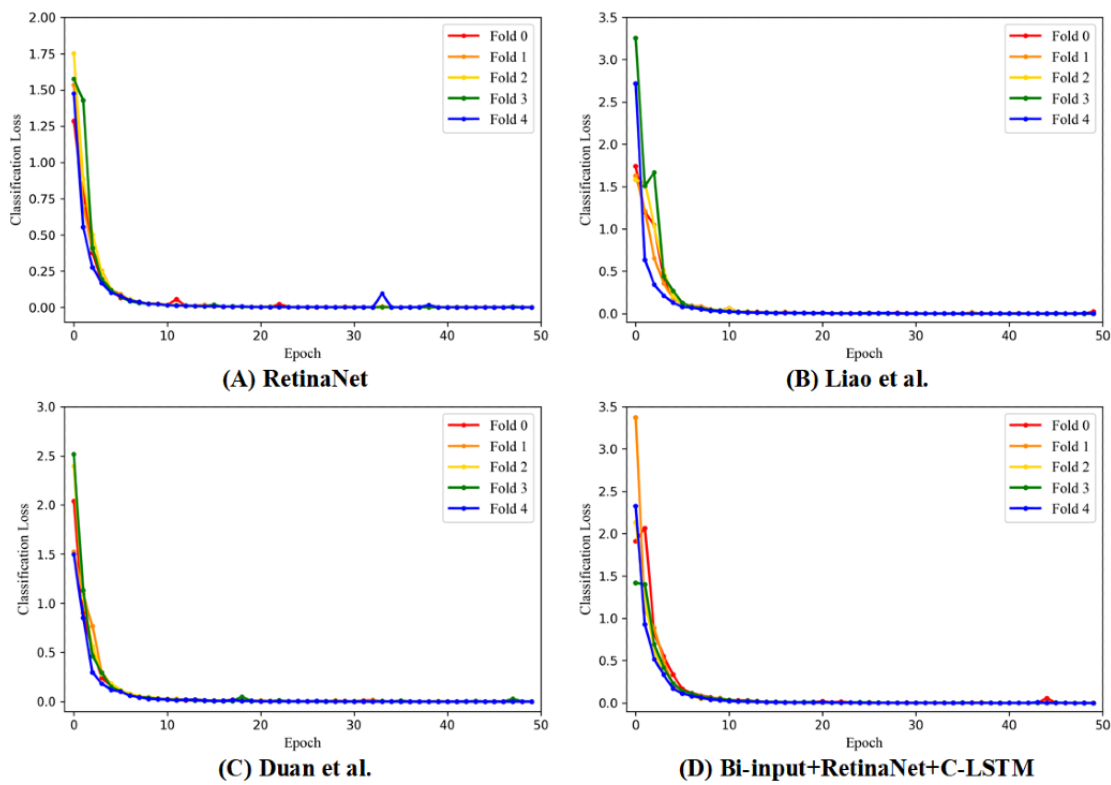
Figure 7. The 5-fold cross-validation results of the smooth L1 loss of each framework. Different color curves indicate different cross-validation models. A: RetinaNet [25]. B: Liao et al [15]. C: Duan et al [16]. D: Bi-input+RetinaNet+C-LSTM. C-LSTM: convolutional long short-term memory.
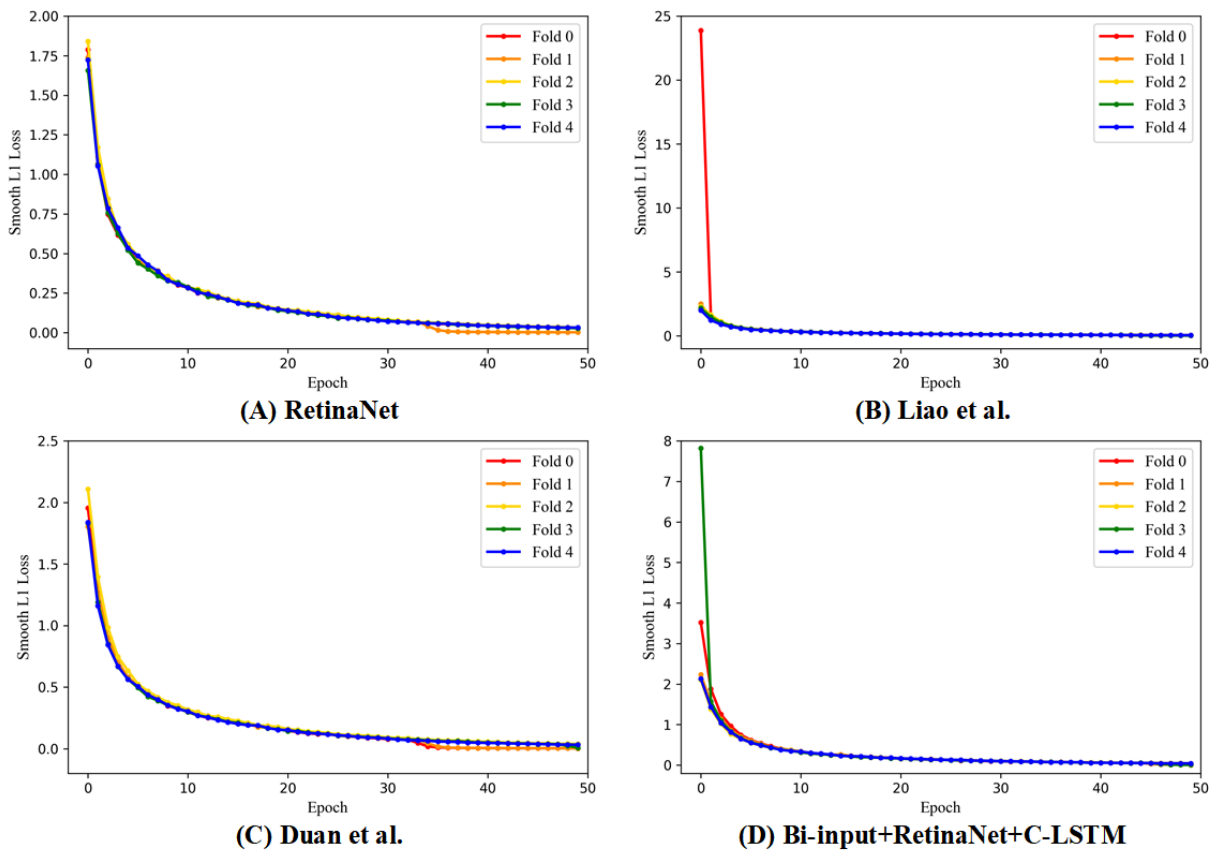
Table 1. The mean average precision (mAP) values from the 5-fold cross-validation.

Frameworks	Fold 1, mAP	Fold 2, mAP	Fold 3, mAP	Fold 4, mAP	Fold 5, mAP
RetinaNet [25]	0.4006	0.6553	0.5687	0.6941	0.7569
Liao et al [15]	0.5082	0.6968	0.5852	0.6479	0.7681
Duan et al [16]	0.4982	0.7157	0.4666	0.7925	0.8294
Bi-input+RetinaNet+C-LSTM[a]	0.4435	0.6523	0.5254	0.6506	0.7408

[a]C-LSTM: convolutional long short-term memory.

Table 2. The performance of each framework.

Frameworks	Sensitivity (%), mean (range)	Specificity (%), mean (range)	Accuracy (%), mean (range)	Time cost (s)
RetinaNet [25]	89 (67.02-98.43)	80 (56.34-94.27)	84.50 (69.57-93.97)	0.24
Liao et al [15]	88 (65.76-98.06)	89 (67.02-98.43)	88.50 (74.44-96.39)	2.21
Duan et al [16]	87 (64.53-97.66)	86 (63.31-97.24)	86.50 (71.97-95.22)	0.33
Bi-input+RetinaNet+C-LSTM[a]	89 (67.02-98.43)	93 (72.30-99.56)	91 (77.63-97.72)	2.72
Human experts	90 (68.30-98.77)	90 (68.30-98.77)	90 (76.34-97.21)	N/A[b]

[a]C-LSTM: convolutional long short-term memory.

[b]N/A: not applicable.

Figure 8. The results of the confusion matrix for each framework. The upper left corners represent true positives, the upper right corners represent false negatives, the lower left corners represent false positives, and the lower right corners represent true negatives. A: RetinaNet [25]. B: Liao et al [15]. C: Duan et al [16]. D: Bi-input+RetinaNet+C-LSTM. C-LSTM: convolutional long short-term memory; Diag+: diagnosed with tumor; Diag-: diagnosed without tumor; Pred+: predicted tumor; Pred-: no predicted tumor.

XSL•FO

RenderX
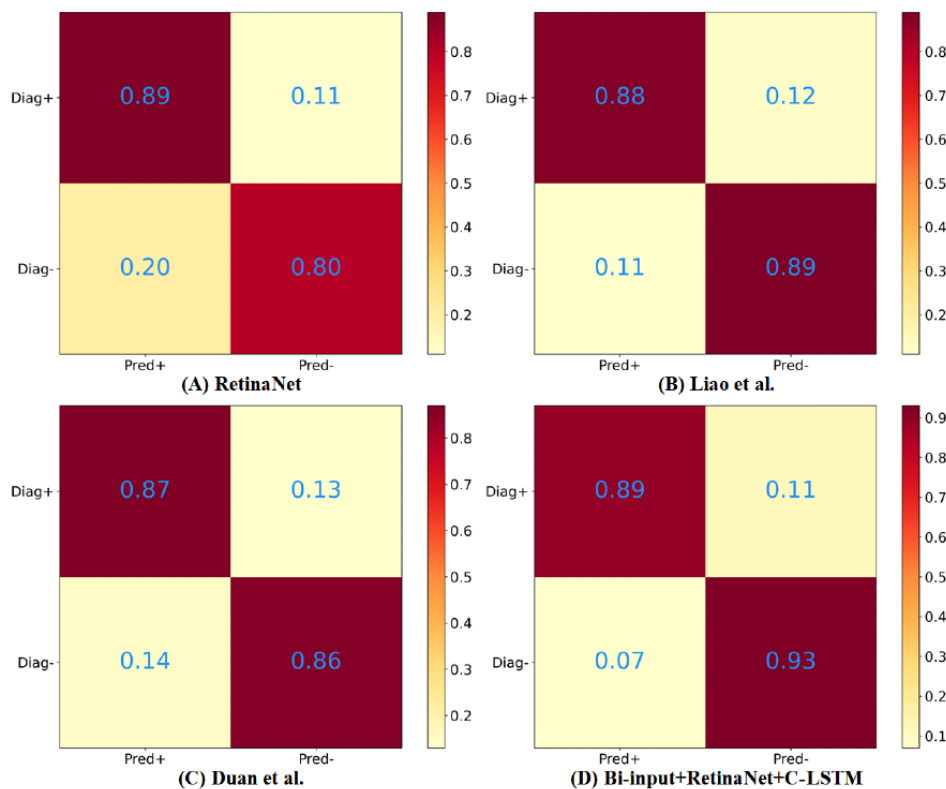
Figure 9. A sample of the original images of the DSA sequence and their corresponding results in the RLS and IADS. A, D, and G represent the raw DSA sequences, and B, E, and H represent the experimental results of the RLS. The results of the IADS are shown in C, F, and I. The red bounding boxes denote the aneurysms, and the green bounding boxes represent the overlapping blood vessels. DSA: digital subtraction angiography; IADS: intracranial aneurysm detection stage; RLS: region localization stage.
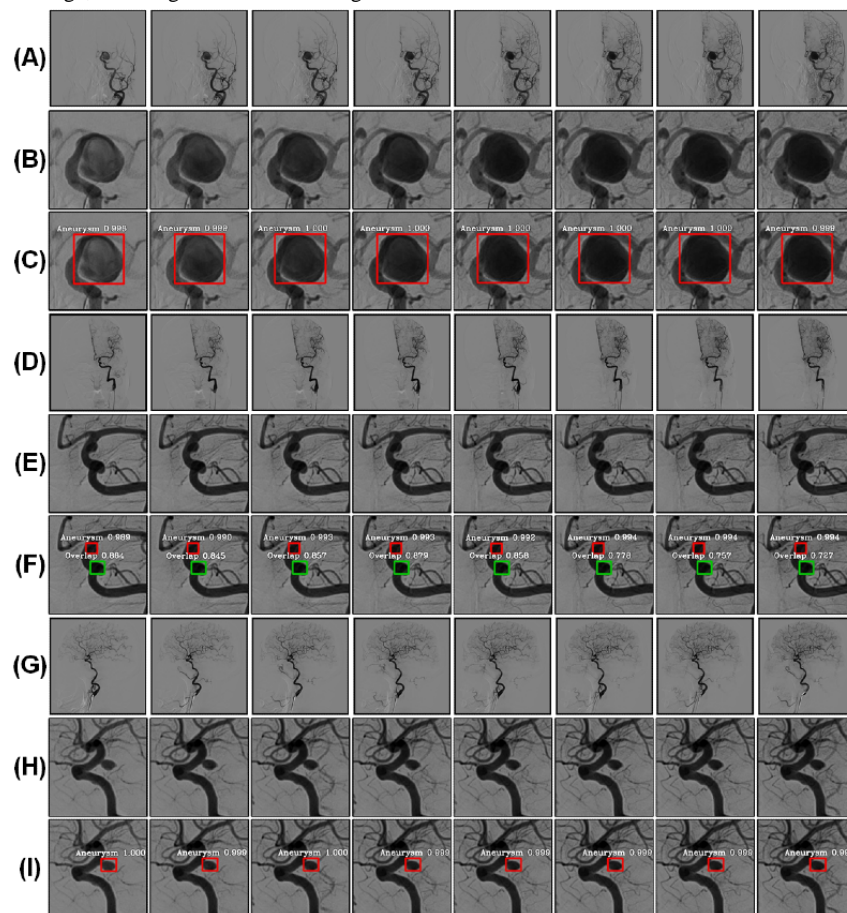
Discussion

Principal Findings

We used the RLS to help decrease the computational load and reduce the interference of unrelated tissues, such as bones and small vessels. This step can reduce time consumption and help neural networks focus on the PCoA region. Moreover, as the 2D DSA images may have had different scales, we used the RLS to standardize the images to the same scale. In the clinical diagnosis process, experienced neurosurgeons and neurointerventional radiologists observed the whole DSA sequence and distinguished overlapping arteries from aneurysms based on the flow of contrast agents through blood vessels. Inspired by this process, we introduced temporal information processing, which has been widely used in text understanding, to improve our diagnostic system. As classic time-processing neural networks, such as LSTM networks, only focus on 1D information, they inevitably result in information loss (ie, the loss of spatial details) when a 2D image is flattened to 1D information. To address this problem, we chose the C-LSTM network, which is specifically designed for 3D data. C-LSTM networks use 3D data as input to process 2D image sequences combined with temporal information. Monodirectional processing methods only allow later features to obtain information from previously inputted images, which results in the imbalance of information. As such, it is difficult to specify

which frame might be more important for detection. Bidirectional temporal information processing allows each frame in DSA sequences to combine both past and future information, and each frame can apply the same weight in the diagnosis process. Although processing time information increases detection times, accuracy is more important than speed when it comes to medical imaging tasks. Even if the detection time increases, the model can still complete the detection within 3 seconds, which is acceptable. Therefore, it was reasonable for us to add a bidirectional C-LSTM network to process information.

In the real diagnosis process, physicians often combine the frontal and lateral sequences to make decisions because some aneurysms are difficult to identify in images taken from 1 angle. Based on this idea, we combined the frontal sequences with the lateral sequences together (bi-input) to increase the amount of spatial information and further improve the performance of the diagnostic system. According to the results of this study, the bi-input+RetinaNet+C-LSTM framework improved the sensitivity to 89% and the specificity to 93%, and its accuracy was the highest (91%) among all models. In addition, the bi-input+RetinaNet+C-LSTM framework also had the highest average AUC value and the best confusion matrix. Hence, the bi-input+RetinaNet+C-LSTM framework had the best performance among all models, and its results were similar to those of experienced human experts.

We labeled some overlapping blood vessels that were easily confused with aneurysms, which also indirectly reduced the rate of false positives to some extent. However, adding the overlap labels also caused fluctuations in the mAP values. The reason for this may have been that the physicians only labeled aneurysms and some overlaps, such as the segment of the ICA near the clinoid process. It was difficult to label all of the overlaps, since our main task was to look for aneurysms, and labeling overlaps requires considerable amounts of work. In our framework's predictions, some overlapping blood vessels were identified by the framework but may not have been marked, and some overlaps were annotated but not detected, which resulted in a large fluctuation in mAP values.

Conclusion

According to our results, more spatial and temporal information can help improve the performance of the frameworks. Therefore,

the bi-input+RetinaNet+C-LSTM had the best performance when compared to that of the other frameworks. Our study demonstrated that our system can assist physicians in detecting intracranial aneurysms on 2D DSA images.

Our experiment had some limitations. First, our data set is comparatively small and only includes PCoA aneurysms. In the future, we will include cerebral aneurysms in different locations. Second, the cascading network framework is relatively complex. Therefore, an end-to-end network should be considered. In future work, we will attempt to find a method that compensates for the loss of information in the process of converting 2D information to 1D information and use a transformer [34] to process time information.

Conflicts of Interest

None declared.

References

1. Wong JHY, Tymianski R, Radovanovic I, Tymianski M. Minimally invasive microsurgery for cerebral aneurysms. Stroke 2015 Sep;46(9):2699-2706. [doi: 10.1161/STROKEAHA.115.008221] [Medline: 26304867]
2. Wardlaw JM, White PM. The detection and management of unruptured intracranial aneurysms. Brain 2000 Feb;123 (Pt 2):205-221. [doi: 10.1093/brain/123.2.205] [Medline: 10648430]
3. Keedy A. An overview of intracranial aneurysms. Mcgill J Med 2020 Dec 01;9(2):141-146 [FREE Full text] [doi: 10.26443/mjm.v9i2.672]
4. Esteva A, Kuprel B, Novoa RA, Ko J, Swetter SM, Blau HM, et al. Dermatologist-level classification of skin cancer with deep neural networks. Nature 2017 Feb 02;542(7639):115-118 [FREE Full text] [doi: 10.1038/nature21056] [Medline: 28117445]
5. Krizhevsky A, Sutskever I, Hinton GE. ImageNet classification with deep convolutional neural networks. 2012 Presented at: Advances in Neural Information Processing Systems 25: 26th Annual Conference on Neural Information Processing Systems 2012; December 3-6, 2012; Lake Tahoe, Nevada, United States p. 1-9 URL: https://papers.nips.cc/paper/2012/file/c399862d3b9d6b76c8436e924a68c45b-Paper.pdf
6. Hubel DH, Wiesel TN. Receptive fields and functional architecture of monkey striate cortex. J Physiol 1968 Mar;195(1):215-243 [FREE Full text] [doi: 10.1113/jphysiol.1968.sp008455] [Medline: 4966457]
7. Fukushima K, Miyake S. Neocognitron: A self-organizing neural network model for a mechanism of visual pattern recognition. In: Competition and Cooperation in Neural Nets. 1982 Presented at: U.S.-Japan Joint Seminar; February 15-19, 1982; Kyoto, Japan p. 267-285. [doi: 10.1007/978-3-642-46466-9_18]
8. Lakhani P, Sundaram B. Deep learning at chest radiography: Automated classification of pulmonary tuberculosis by using convolutional neural networks. Radiology 2017 Aug;284(2):574-582. [doi: 10.1148/radiol.2017162326] [Medline: 28436741]
9. Yasaka K, Akai H, Abe O, Kiryu S. Deep learning with convolutional neural network for differentiation of liver masses at dynamic contrast-enhanced CT: A preliminary study. Radiology 2018 Mar;286(3):887-896. [doi: 10.1148/radiol.2017170706] [Medline: 29059036]
10. Shi X, Chen Z, Wang H, Yeung DY, Wong WK, Woo WC. Convolutional LSTM network: a machine learning approach for precipitation nowcasting. 2015 Dec Presented at: The 28th International Conference on Neural Information Processing Systems; December 7-12, 2015; Montreal, Canada p. 802-810.
11. Novikov AA, Major D, Wimmer M, Lenis D, Buhler K. Deep sequential segmentation of organs in volumetric medical scans. IEEE Trans Med Imaging 2019 May;38(5):1207-1215. [doi: 10.1109/TMI.2018.2881678] [Medline: 30452352]

XSL·FO

RenderX

12.   Liu X, Liu T, Zhang Z, Kuo PC, Xu H, Yang Z, et al. TOP-Net prediction model using bidirectional long short-term memory and medical-grade wearable multisensor system for tachycardia onset: Algorithm development study. JMIR Med Inform 2021 Apr 15;9(4):e18803 [FREE Full text] [doi: 10.2196/18803] [Medline: 33856350]

13.   Podgoršak AR, Bhurwani MM, Rava RA, Chandra AR, Ionita CN. Use of a convolutional neural network for aneurysm identification in digital subtraction angiography. 2019 Mar 13 Presented at: SPIE Medical Imaging 2019: Computer-Aided Diagnosis; February 16-21, 2019; San Diego, California, United States. [doi: 10.1117/12.2512810]

14.   Jin H, Yin Y, Hu M, Yang G, Qin L. Fully automated unruptured intracranial aneurysm detection and segmentation from digital subtraction angiography series using an end-to-end spatiotemporal deep neural network. 2019 Mar 15 Presented at: SPIE Medical Imaging 2019: Image Processing; February 16-21, 2019; San Diego, California, United States. [doi: 10.1117/12.2512623]

15.   Liao J, Duan H, Dai H, Huang Y, Liu L, Chen L, et al. Automatic detection of intracranial aneurysm from digital subtraction angiography with cascade networks. 2019 Aug Presented at: The 2nd International Conference on Artificial Intelligence and Pattern Recognition; August 16-18, 2019; Beijing, China p. 18-23. [doi: 10.1145/3357254.3357258]

16.   Duan H, Huang Y, Liu L, Dai H, Chen L, Zhou L. Automatic detection on intracranial aneurysm from digital subtraction angiography with cascade convolutional neural networks. Biomed Eng Online 2019 Nov 14;18(1):110 [FREE Full text] [doi: 10.1186/s12938-019-0726-2] [Medline: 31727057]

17.   Ojemann RG, Crowell RM. Surgical management of cerebrovascular disease. Ann Surg 1984;199(3):49A. [doi: 10.1097/00000658-198403000-00022]

18.   Lin TY, Dollár P, Girshick R, He K, Hariharan B, Belongie S. Feature pyramid networks for object detection. 2017 Nov 09 Presented at: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR); July 21-26, 2017; Honolulu, Hawaii, USA. [doi: 10.1109/cvpr.2017.106]

19.   He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition. 2016 Dec 12 Presented at: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR); June 27-30, 2016; Las Vegas, Nevada, USA. [doi: 10.1109/cvpr.2016.90]

20.   Ren S, He K, Girshick R, Sun J. Faster R-CNN: Towards real-time object detection with region proposal networks. IEEE Trans Pattern Anal Mach Intell 2017 Jun;39(6):1137-1149. [doi: 10.1109/TPAMI.2016.2577031] [Medline: 27295650]

21.   Zhang K, Liu X, Liu F, He L, Zhang L, Yang Y, et al. An interpretable and expandable deep learning diagnostic system for multiple ocular diseases: Qualitative study. J Med Internet Res 2018 Nov 14;20(11):e11144 [FREE Full text] [doi: 10.2196/11144] [Medline: 30429111]

22.   Ko H, Chung H, Kang WS, Kim KW, Shin Y, Kang SJ, et al. COVID-19 pneumonia diagnosis using a simple 2D deep learning framework with a single chest CT image: Model development and validation. J Med Internet Res 2020 Jun 29;22(6):e19569 [FREE Full text] [doi: 10.2196/19569] [Medline: 32568730]

23.   Liang B, Yang N, He G, Huang P, Yang Y. Identification of the facial features of patients with cancer: A deep learning–based pilot study. J Med Internet Res 2020 Apr 29;22(4):e17234. [doi: 10.2196/17234]

24.   Huang P, Yu G, Lu H, Liu D, Xing L, Yin Y, et al. Attention-aware fully convolutional neural network with convolutional long short-term memory network for ultrasound-based motion tracking. Med Phys 2019 May;46(5):2275-2285. [doi: 10.1002/mp.13510] [Medline: 30912590]

25.   Lin TY, Goyal P, Girshick R, He K, Dollár P. Focal loss for dense object detection. 2017 Dec 25 Presented at: 2017 IEEE International Conference on Computer Vision (ICCV); October 22-29, 2017; Venice, Italy. [doi: 10.1109/iccv.2017.324]

26.   Long J, Shelhamer E, Darrell T. Fully convolutional networks for semantic segmentation. 2015 Oct 15 Presented at: 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR); June 7-12, 2015; Boston, Massachusetts, USA p. 3431-3440. [doi: 10.1109/cvpr.2015.7298965]

27.   Jung H, Kim B, Lee I, Yoo M, Lee J, Ham S, et al. Detection of masses in mammograms using a one-stage object detector based on a deep convolutional neural network. PLoS One 2018 Sep 18;13(9):e0203355. [doi: 10.1371/journal.pone.0203355] [Medline: 30226841]

28.   Umer J, Irtaza A, Nida N. MACCAI LiTS17 liver tumor segmentation using RetinaNet. 2021 Jan 20 Presented at: 2020 IEEE 23rd International Multitopic Conference (INMIC); November 5-7, 2020; Bahawalpur, Pakistan. [doi: 10.1109/inmic50486.2020.9318116]

29.   Gräbel P, Özkan Ö, Crysandt M, Herwartz R, Baumann M, Klinkhammer BM, et al. Circular anchors for the detection of hematopoietic cells using Retinanet. 2020 May 22 Presented at: 2020 IEEE 17th International Symposium on Biomedical Imaging (ISBI); April 3-7, 2020; Iowa City, Iowa, USA. [doi: 10.1109/isbi45749.2020.9098398]

30.   Keras: The Python Deep Learning library. The SAO/NASA Astrophysics Data System. URL: https://ui.adsabs.harvard.edu/abs/2018ascl.soft06022C/abstract [accessed 2022-03-11]

31.   Kingma D, Ba J. Adam: A method for stochastic optimization. arXiv. Preprint posted online on December 22, 2014 [FREE Full text]

32.   DeLong ER, DeLong DM, Clarke-Pearson DL. Comparing the areas under two or more correlated receiver operating characteristic curves: A nonparametric approach. Biometrics 1988 Sep;44(3):837-845. [doi: 10.2307/2531595]

33.   Agresti A, Coull BA. Approximate is better than "Exact" for interval estimation of binomial proportions. Am Stat 1998;52(2):119-126. [doi: 10.1080/00031305.1998.10480550]

34.   Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, et al. Attention is all you need. arXiv. Preprint posted online on December 6, 2017 [FREE Full text]

Abbreviations

AUC: area under the curve
C-LSTM: convolutional long short-term memory
CNN: convolutional neural network
DSA: digital subtraction angiography
FPN: feature pyramid network
IADS: intracranial aneurysm detection stage
ICA: internal carotid artery
LSTM: long short-term memory
mAP: mean average precision
PCoA: posterior communicating artery
ResNet: residual deep neural network
RLS: region localization stage
ROC: receiver operating characteristic
SAH: subarachnoid hemorrhage

<u>Original Paper</u>

A Bayesian Network Analysis of the Probabilistic Relationships Between Various Obesity Phenotypes and Cardiovascular Disease Risk in Chinese Adults: Chinese Population-Based Observational Study

Simiao Tian[1], PhD; Mei Bi[2], MM; Yanhong Bi[1], MPH; Xiaoyu Che[1], MPH; Yazhuo Liu[2], MM

[1]Department of Research, Affiliated Zhongshan Hospital of Dalian University, Dalian, China

[2]Department of Clinical Nutrition and Metabolism, Affiliated Zhongshan Hospital of Dalian University, Dalian, China

Corresponding Author:
Simiao Tian, PhD
Department of Research
Affiliated Zhongshan Hospital of Dalian University
No. 6, Jiefang street
Dalian, 116001
China
Phone: 86 41162898583
Email: tiansimiao@dlu.edu.cn

Abstract

Background: Cardiovascular disease (CVD) risk among individuals with different BMI levels might depend on their metabolic health. The extent to which metabolic health status and BMI affect CVD risk, either directly or through a mediator, in the Chinese population remains unclear.

Objective: In this study, the Bayesian network (BN) perspective is adopted to characterize the multivariable probabilistic connections between CVD risk and metabolic health and obesity status and identify potential factors that influence these relationships among Chinese adults.

Methods: The study population comprised 6276 Chinese adults aged 30 to 74 years who participated in the China Health and Nutrition Survey 2009. BMI was used to categorize participants as normal weight, overweight, or obese, and metabolic health was defined by the Adult Treatment Panel-3 criteria. Participants were categorized into 6 phenotypes according to their metabolic health and BMI categorization. The 10-year risk of CVD was determined using the Framingham Risk Score. BN modeling was used to identify the network structure of the variables and compute the conditional probability of CVD risk for the different metabolic obesity phenotypes with the given structure.

Results: Of 6276 participants, 64.67% (n=4059), 20.37% (n=1279), and 14.95% (n=938) had a low, moderate, and high 10-year CVD risk. An averaged BN with a stable network structure was constructed by learning 300 bootstrapped networks from the data. Using BN reasoning, the conditional probability of high CVD risk increased as age progressed. The conditional probability of high CVD risk was 0.43% (95% CI 0.2%-0.87%) for the 30 to 40 years age group, 2.25% (95% CI 1.75%-2.88%) for the 40 to 50 years age group, 16.13% (95% CI 14.86%-17.5%) for the 50 to 60 years age group, and 52.02% (95% CI 47.62%-56.38%) for those aged ≥70 years. When metabolic health and BMI categories were instantiated to their different statuses, the conditional probability of high CVD risk increased from 7.01% (95% CI 6.27%-7.83%) for participants who were metabolically healthy normal weight to 10.47% (95% CI 7.63%-14.18%) for their metabolically healthy obese (MHO) counterparts and up to 21.74% and 34.48% among participants who were metabolically unhealthy normal weight and metabolically unhealthy obese (MUO), respectively. Sex was a significant modifier of the conditional probability distribution of metabolic obesity phenotypes and high CVD risk, with a conditional probability of high CVD risk of only 2.02% and 22.7% among MHO and MUO women, respectively, compared with 21.92% and 48.21% for their male MHO and MUO counterparts, respectively.

Conclusions: BN modeling was applied to investigate the relationship between CVD risk and metabolic health and obesity phenotypes in Chinese adults. The results suggest that both metabolic health and obesity status are important for CVD prevention; closer attention should be paid to BMI and metabolic status changes over time.

XSL•FO

RenderX

Introduction

Background

Cardiovascular disease (CVD) is becoming a leading cause of mortality, disability, and rising health care costs worldwide [1,2]. The worldwide prevalence of CVD doubled from 271 million in 1990 to 523 million in 2019 [1], and recent epidemiological studies indicate that CVD accounts for >40% of deaths in the general Chinese population [3]. Despite significant efforts directed toward CVD prevention and control at the individual level and public health level, there has still been a clear increase in deaths because of CVD in China over the past 2 decades, from 2.51 million in 1990 to 3.97 million in 2016, as well as a doubled prevalence from 1990 to 2016 [2,3].

Obesity is recognized as the primary cause of many chronic diseases. It is also an established risk factor for CVDs, including coronary disease [4], myocardial infarction [5], and ischemic heart disease [6]. Previous cohort studies have reported a causal relationship between obesity and increased risk of CVD mortality [7]. Along with obesity, metabolic syndrome (MetS), which is a cluster of interrelated metabolic abnormalities, including increased blood pressure (BP), hyperglycemia, central adiposity, insulin resistance, and dyslipidemia, is another well-established determinant of CVD and mortality [8]. However, there is heterogeneity in body fat distribution and metabolic factors among individuals with obesity, and it has been reported that a subgroup of people with obesity possesses a favorable cardiometabolic profile; these individuals are referred to as people who are metabolically healthy (MH) obese. They may not be at increased risk of several health outcomes, including CVD [9], and may even confer a protective effect on all-cause mortality if accompanied by a healthy metabolism [10]. Together, the findings of these studies highlight the need to take metabolic health and obesity status into account in CVD-related studies.

In general, any condition or disease that affects the heart, its vessels, and the blood circulatory system [11] or is associated with conditions such as chronic heart failure (HF), congenital heart disease, rhythm disorders, and subclinical atherosclerosis [12] can be related to CVD. In addition to the main risk factors, recently published studies have highlighted the important role of other factors such as infection, inflammatory conditions, and chronic diseases in CVD development [13]. CVD is a multicausal disease and presents a clear heterogeneity in terms of prevalence and mortality among various subgroups that differ in their demographic characteristics; therefore, sex, age, smoking, high cholesterol, hypertension, and diabetes should be taken into account in CVD studies [14], together with metabolic health and BMI levels. Modeling multiple correlated factors when assessing CVD risk can be computationally challenging and requires new statistical approaches. Standard regression modeling requires independence among covariates and cannot disentangle the interrelationships or interactions that form complex networks of relationships. Bayesian networks (BNs) are powerful probabilistic graphical models that enable the description of conditional dependencies and reasoning among a set of variables, permitting a comprehensive investigation of interrelationships among multiple correlated variables and identification of potential causality [15]. The generated BN model can be used for dynamic qualitative and quantitative reasoning, where the probability of all variables changes by updating the state of one variable, revealing inferences between the depicted variables [16]. Recently, BNs have been extensively used in health science and epidemiology, particularly in CVD research in areas such as diagnosis, risk assessment, and disease prediction [17-19]. To our knowledge, few studies have examined the interrelationships between metabolic health status, BMI, and CVD risk in conjunction with demographic factors, biomarkers, and chronic health outcomes in the Chinese population.

Objective

This study aims to fill this gap in knowledge by introducing BN modeling and evaluating the multivariable probabilistic connections among metabolic health, obesity, and CVD risk in a population-level study of Chinese adults. In addition, this study aims to identify factors that directly and indirectly influence these relationships.

Methods

Study Population

The participants in this study were recruited from the China Health and Nutrition Survey (CHNS), which is an ongoing longitudinal survey designed to examine the effects of health and nutrition at the population level. A detailed description of the CHNS, such as the multistage sampling design and data collection methods, has been provided elsewhere [20], and this study uses a cohort that has been previously described [21]. Briefly, the participants included in this study were obtained from the 2009 CHNS wave (N=11,929). The participants voluntarily participated in health interviews and examinations, answered the general sociodemographic questions, and completed an in-depth health questionnaire. Data were collected via household interviews.

Ethics Approval and Consent to Participate

The CHNS study was approved by the institutional review committees of the University of North Carolina at Chapel Hill, the National Institute of Nutrition and Food Safety, Chinese Centers for Disease Control and Prevention, the China–Japan Friendship Hospital, and the Ministry of Health (R01-HD30880, DK056350, and R01-HD38700). All participants provided written informed consent. All methods were performed in accordance with the relevant guidelines and regulations.

XSL•FO

RenderX

Data Collection and Measurements

Face-to-face interviews were conducted by well-trained personnel using self-administered and standardized questionnaires. The interviews were used to collect information on participants' demographic characteristics (age, sex, marital status, and education level), behavioral factors (smoking status, drinking status, and physical activity), medication use, and self-reported family history.

All participants underwent a physical examination performed by well-trained examiners, following standardized procedures. Body weight and height were measured to the nearest 0.1 kg and 0.1 cm, respectively, with the participants wearing light clothing and no shoes. Waist circumference was measured to the nearest 0.1 cm at the midpoint between the bottom of the rib cage and the top of the iliac crest following exhalation. BMI was calculated as weight (kg) divided by the square of height (meters). Systolic BP (SBP) and diastolic BP (DBP) were measured using a standardized mercury sphygmomanometer on the participant's right arm. BP measurements were performed in triplicate after 10 minutes of seated rest, and the mean of the 3 measurements was used in the analyses.

The participants were required to fast overnight before blood collection. Fasting blood samples were obtained the following morning using a standardized process and were then analyzed in a national central clinical laboratory in Beijing. Plasma and serum samples were frozen and stored at −86 °C for later laboratory analyses. Serum levels of fasting plasma glucose (FPG), total cholesterol (TC), low-density lipoprotein cholesterol, high-density lipoprotein cholesterol (HDL-C), triglyceride (TG), uric acid (UA), and other routine blood biochemical indices were measured using a biochemical autoanalyzer. Details of all laboratory analyses and measurements can be found elsewhere [20]. Homeostasis model assessment of insulin resistance (HOMA-IR) was calculated using the following formula:

$$\text{HOMA-IR} = \text{fasting insulin (microinternational units per milliliter)} \times \text{FPG (millimoles per liter)}/22.5 \quad (1)$$

Fasting serum was used to derive the serum creatinine concentration (mg/dL). The estimated glomerular filtration rate (eGFR) was calculated using the Chronic Kidney Disease Epidemiology equations combined with the serum creatinine equation. The robust performance of serum creatinine–based equations has been validated in the Chinese population [22].

In this study, according to the criteria recommended by the US Joint National Committee and Chinese guidelines [23,24], hypertension was defined as an SBP ≥140 mm Hg, a DBP ≥90 mm Hg, and/or the self-reported use of antihypertensive medication. Diabetes was defined as FPG ≥7.0 mmol/L or treatment for diabetes. On the basis of the National Cholesterol Education Project guidelines [25], dyslipidemia was defined as low-density lipoprotein cholesterol ≥4.14 mmol/L, HDL-C ≤1.036 mmol/L, and TGs ≥2.26 mmol/L. Hyperuricemia was defined as serum UA ≥420 μmol/L in men and ≥360 μmol/L in women [26].

Assessment of 10-Year Risk of CVD

The Framingham Risk Score (FRS) was used to estimate the 10-year probability of a CVD event (coronary heart disease, cerebrovascular event, peripheral artery disease, or HF). The FRS was developed by D'Agostino et al [27] using a sex-specific multivariable risk factor algorithm and has been validated in American, Canadian, European, and Asian populations [28-31], as well as in Chinese participants [32]. It is used in primary care to assess overall cardiovascular risk among participants who are asymptomatic at baseline and are aged 30 to 74 years, providing clinicians with quantitative information to aid in the targeted lowering of risk factors [33]. As per the conditions of the FRS algorithm, participants s aged <30 years or >74 years, as well as those with incomplete data with respect to the anthropometric measures and blood sampling, were excluded. As a result, a total of 6276 individuals (2895, 46.13%) men and (3381, 53.87%) women were enrolled in this study.

The raw FRS score was calculated for each participant based on the individual's sex, age, TC, smoking status, HDL-C, SBP (treatment for hypertension and SB*P* value), and diabetes status, together with their associated proper β coefficient value from the proportional hazard regression [27]. The 10-year risk factor was then derived as a percentage by gender. In addition, the 10-year CVD risk was categorized as low (FRS <10%), moderate (10%-19%), or high (≥20%) according to previous recommendations [13,28].

Definitions of Obesity, Metabolic Health, and Metabolic Obesity Phenotypes

Overweight and obesity were defined as BMI≥24 kg/m² and ≥28 kg/m², respectively, using the criteria for Chinese adults [34,35]. Each participant was then categorized into one of three BMI groups: normal weight (BMI 18.5-23.9 kg/m²), overweight (BMI 24.0-27.9 kg/m²), and obese (BMI ≥28.0 kg/m²).

The metabolic health status of each participant was defined based on the Adult Treatment Panel-3 definition of MetS [36]. Participants who met ≥2 of the following four criteria were considered metabolically unhealthy (MU): (1) hypertension (SBP/DBP ≥130/85 mm Hg or use of antihypertensive drugs), (2) hypertriglyceridemia (TG ≥1.7 mmol/L or use of lipid-lowering drugs), (3) hyperglycemia (FPG ≥5.6 mmol/L or use of medications for diabetes), and (4) reduced HDL-C (HDL-C <1.04 mmol/L for men and <1.3 mmol/L for women). The waist circumference criterion was not used because of collinearity with BMI.

The above definition was used together with the BMI categories to classify the study participants into one of six following metabolic obesity phenotypes: participants who are MH participants with normal weight, participants who are MH and overweight, participants who are MH and obese (MH normal weight [MHNW], MH overweight [MHOW], and MH obese [MHO], respectively), participants who are MU with normal weight, participants who are MU and overweight, and participants who are MU and obese (MU normal weight [MUNW], MU overweight [MUOW], and MU obese [MUO], respectively).

BN Modeling

A BN is a probabilistic graphical model that represents a set of random variables $X = \{X_i,...,X_n\}$ as nodes and their conditional dependencies as edges through a directed acyclic graph (DAG) [16]. A BN can be fully specified by a pair (G, P), in which G=(V, A) is a DAG comprising nodes (denoted by V) and directed edges (denoted by A) and P is a joint probability distribution. Specifically, if there is an edge from node X_i to node X_j, X_i is then termed the parent and X_j, the child, and the direction of the edge indicates a statistical dependence between the corresponding variables. The joint distribution P can be written as the product of the local conditional probability of each node X_i, given its parent variables in graph G, as follows:

$Pa(X_i)$ are the parents of X_i in the BN, and P(X) reflects the properties of the BN.

The BN models were built and reviewed through an iterative 2-stage process, including a stepwise manual construction process and a data-driven approach [37,38]. First, a manual construction approach was used to explore different network structures by including various sets of potential risk factors or variables consecutively. The selection of the variable nodes was based on prior expert knowledge and a systematic review of the literature, which has been shown to improve BN structure learning processes and to avoid excessive complexity of the network structure by the inclusion of too many nodes [39]. During this first stage, prior knowledge can be included in the model as blacklist and whitelist arcs. Specifically, the directions between certain variables were restricted by using a layering approach [40]. For instance, the variables metabolic health status and BMI were allowed to be directed to FRS categories, and this setting ensured that information was embedded in the direction of causality for the effect of different obesity phenotypes (whitelist), and the FRS categories were not permitted to influence age (blacklist), as we were interested in understanding the age-related pathways that explain 10-year CVD risk as an outcome.

Second, a data-driven approach using different structure learning algorithms was adopted to further improve the BN. On the basis of prespecified simulations and comparison among score-based, constraint-based, and hybrid structure learning approaches, the tabu-search algorithm [41] was used for graphical structure learning along with the Bayesian information criterion score [42] to achieve high quality of the network structure. The stabilities of the arcs in the network were examined from 300 bootstrapped networks, and the arc strengths (between 0 and 1) were estimated by averaging the probability of the arcs presenting in these bootstrap-resampled network structures [21,43]. The final BN model was obtained by using the structure and directions of arcs from the averaged network and was then

further used to query the conditional probability distributions (Bayesian reasoning) with a specific value or evidence provided.

Statistical Analyses

Continuous variables were presented as medians or means with SD according to its assumption of normality from the Kolmogorov–Smirnov test, and categorical variables were presented as numerical variables with the corresponding proportions as functions of metabolic health status and BMI. Comparison of the different obesity phenotypes was performed using 1-way analysis of variance or the Kruskal-Wallis test for continuous variables, where appropriate, or the chi-square test for categorical variables. P values for trend were computed using Pearson for continuous variables and the Mantel–Haenszel chi-square test for categorical variables. All statistical analyses were performed using R (version 3.2.2; R Foundation for Statistical Computing) software [44], and $P<.05$ was considered statistically significant. The bnlearn package in the R software environment was used for BN modeling analysis [43], including network structure learning, parameter estimation, network arc stabilities, conditional probability queries in the finalized network, and visualization. The data and code for full analysis can be obtained by reasonable request from the corresponding author.

Results

Characteristics of the Sample

The characteristics of the sample, stratified by BMI and metabolic health status, are shown in Table 1. Among the 3881 participants without MetS, 2548 (65.65%) had a normal BMI, and 198 (5.1%) had MHO. Among 2395 participants with MU profiles, 955 (39.87%), 981 (40.96%), and 459 (19.16%) were classified into the MUNW, MUOW, and MUO groups, respectively. Within the same BMI levels, the MU groups more commonly exhibited greater waist and hip circumference measurements, along with elevated BP, TC, TG, and UA, and lower levels of HDL-C than the MH groups. In particular, glucose biomarkers, including FPG, HOMA-IR, and Hemoglobin A_{1c}, were higher in the participants who were MU than in their healthy counterparts (Table 1). The distributions of age groups, sex, smoking, and alcohol drinking status also differed among the 6 groups ($P<.001$, $P=.002$, $P<.001$, and $P=.01$, respectively). Participants who were MHO were more often women and were more commonly nonsmokers and nondrinkers compared with their obese counterparts with MetS (MUO), whereas a more unfavorable risk profile was seen in the MUO group than in the MHO group. Among the 955 participants in the MUNW group, 419 (43.9%) had hypertension, 112 (11.7%) had diabetes, and 545 (57.1%) had dyslipidemia. Among the 459 participants in the MUO group, the prevalence of the abovementioned cardiometabolic disorders was 279 (60.8%), 74 (16.1%), and 338 (73.6%), respectively.

Table 1. Characteristics of study sample based on combinations of BMI and metabolic health defined by Adult Treatment Panel-3 criteria (N=6276).

Characteristics	MHNW[a] (n=2548)	MHOW[b] (n=1135)	MHO[c] (n=198)	MUNW[d] (n=955)	MUOW[e] (n=981)	MUO[f] (n=459)	P value[g]	P value for trend[h]
Age (years), mean (SD)	49.53 (11.21)	50.43 (10.66)	49.68 (10.88)	54.66 (10.77)	53.15 (10.30)	52.36 (10.68)	<.001	<.001
Sex (male), n (%)	1188 (46.6)	507 (44.7)	71 (35.9)	432 (45.2)	496 (50.6)	201 (43.8)	.002	.59
Smoker, n (%)	859 (33.7)	314 (27.7)	37 (18.7)	315 (33)	341 (34.8)	125 (27.2)	<.001	.49
Alcohol drinker, n (%)	882 (34.6)	398 (35.1)	59 (29.8)	290 (30.4)	368 (37.5)	144 (31.4)	.01	.52
Weight (kg), mean (SD)	55.72 (6.83)	66.66 (7.14)	76.08 (9.27)	56.99 (7.18)	67.73 (7.82)	78.42 (9.45)	<.001	<.001
Height (cm), mean (SD)	160.86 (8.12)	161.25 (8.04)	160.10 (8.92)	160.70 (8.50)	161.83 (8.57)	161.00 (8.61)	.01	.12
BMI (kg/m^2), mean (SD)	21.48 (1.44)	25.58 (1.10)	29.59 (1.51)	22.00 (1.40)	25.79 (1.12)	30.18 (1.92)	<.001	<.001
Waist circumference (cm), mean (SD)	77.76 (7.12)	87.30 (6.82)	95.40 (8.25)	80.70 (7.43)	89.06 (6.72)	98.27 (7.50)	<.001	<.001
Hip circumference (cm), mean (SD)	91.16 (5.47)	98.27 (5.75)	105.01 (6.14)	92.16 (6.24)	98.59 (5.39)	106.07 (6.15)	<.001	<.001
HDL-C[i] (mmol/L), mean (SD)	1.58 (0.49)	1.46 (0.34)	1.47 (0.34)	1.31 (0.39)	1.20 (0.33)	1.18 (0.60)	<.001	<.001
LDL-C[j] (mmol/L), mean (SD)	2.93 (0.91)	3.16 (0.84)	3.27 (0.88)	3.01 (1.07)	3.10 (1.05)	3.13 (1.23)	<.001	<.001
DBP[k] (mm Hg), mean (SD)	76.98 (9.78)	80.54 (10.45)	83.17 (9.92)	84.06 (11.11)	86.21 (10.44)	89.75 (12.11)	<.001	<.001
SBP[l] (mm Hg), mean (SD)	118.08 (15.30)	123.11 (16.81)	127.00 (14.80)	130.66 (18.46)	133.19 (18.00)	137.72 (20.37)	<.001	<.001
FPG[m] (mmol/L), mean (SD)	4.94 (0.66)	5.04 (0.79)	5.13 (0.94)	5.91 (1.54)	5.92 (1.71)	6.11 (1.67)	<.001	<.001
TC[n] (mmol/L), mean (SD)	4.71 (0.91)	4.90 (0.91)	5.00 (0.92)	5.04 (1.08)	5.21 (1.04)	5.22 (1.07)	<.001	<.001
TG[o] (mmol/L), mean (SD)	1.09 (0.65)	1.27 (0.70)	1.25 (0.51)	2.33 (1.55)	2.82 (2.10)	2.96 (2.15)	<.001	<.001
Urea (mmol/L), mean (SD)	5.43 (1.63)	5.45 (1.45)	5.33 (1.19)	5.54 (1.45)	5.65 (1.48)	5.59 (1.47)	.001	<.001
Uric acid (μmol/L), mean (SD)	276.75 (82.64)	291.26 (84.42)	292.03 (75.75)	325.58 (106.51)	358.75 (131.02)	362.59 (115.99)	<.001	<.001
HOMA-IR[p], mean (SD)	2.38 (3.15)	2.93 (4.40)	3.43 (2.74)	4.89 (11.37)	5.23 (7.93)	6.26 (6.58)	<.001	<.001
hsCRP[q], mean (SD)	1.81 (5.24)	2.21 (5.25)	2.77 (4.40)	2.82 (10.08)	2.94 (5.66)	3.68 (5.22)	<.001	<.001
HbA$_{1C}$[r], mean (SD)	5.41 (0.53)	5.54 (0.53)	5.66 (0.69)	5.67 (0.94)	5.83 (0.94)	6.05 (1.02)	<.001	<.001
Age groups, n (%)								
30-39	609 (23.9)	225 (19.8)	43 (21.7)	102 (10.7)	119 (12.1)	63 (14)	<.001	<.001
40-49	738 (29)	349 (30.7)	63 (31.8)	220 (23)	252 (25.7)	139 (30.2)	<.001	<.001
50-59	700 (27.5)	322 (28.4)	52 (26.3)	311 (32.6)	346 (35.3)	149 (32.5)	<.001	<.001
60-69	390 (15.3)	194 (17.1)	33 (16.7)	246 (25.8)	212 (21.6)	78 (17)	<.001	<.001
≥70	111 (4.4)	45 (4)	7 (3.5)	76 (8)	52 (5)	30 (7)	<.001	<.001
Hypertension, n (%)	379 (14.9)	251 (22.1)	70 (35.4)	419 (43.9)	511 (52.1)	279 (60.8)	<.001	<.001
Diabetes, n (%)	21 (0.8)	10 (0.9)	3 (1.5)	112 (11.7)	110 (11.2)	74 (16)	<.001	<.001
Dyslipidemia, n (%)	372 (14.6)	234 (20.6)	46 (23.2)	545 (57.1)	696 (71)	338 (73.6)	<.001	<.001
Hyperuricemia, n (%)	170 (6.7)	119 (10.5)	18 (9.1)	187 (19.6)	302 (30.8)	156 (34)	<.001	<.001

[a]MHNW: metabolically healthy normal weight.

[b]MHOW: metabolically healthy overweight.

[c]MHO: metabolically healthy obese.

[d]MUNW: metabolically unhealthy normal weight.

[e]MUOW: metabolically unhealthy overweight.

[f]MUO: metabolically unhealthy obese.

[g]The *P* value for overall comparison of the different obesity phenotypes.

[h]The *P* value for trend was computed from the Pearson test for continuous variables and the Mantel-Haenszel chi-square test for categorical variables.

[i]HDL-C: high-density lipoprotein cholesterol.

[j]LDL-C: low-density lipoprotein cholesterol.

[k]DBP: diastolic blood pressure.

[l]SBP: systolic blood pressure.

[m]FPG: fasting plasma glucose.

[n]TC: total cholesterol.

[o]TG: triglyceride.

[p]HOMA-IR: homeostatic model assessment of insulin resistance.

[q]hsCRP: high-sensitivity C-reactive protein.

[r]HbA_{1c}: hemoglobin A_{1c}.

Distribution of FRS According to Obesity Phenotypes

The distribution of FRS according to metabolic health status and BMI is shown in Table 2. In general, the FRS increased with weight in both participants who were MH and participants who were MU. The average FRS among participants who were obese and with favorable metabolic profiles (MHO) was 7.54% (SD 7.91%), whereas the risk score doubled (14.16%, SD 13.01%) among participants who were MUNW and further increased to 15.98% (SD 14.42%) among participants who were MUO (Table 2).

With regard to FRS categories, among those with a MU profile, the proportion of participants with a high 10-year CVD risk

ranged from 24% (229/955) in the MUNW group to 27.5% (125/459) in the MUO group. In addition, approximately half of all participants who were MU had a low 10-year CVD risk, with proportions of 50.2% (497/955), 46.1% (452/981), and 45.8% (210/459) among the MUNW, MUOW, and MUO groups, respectively. In contrast, among participants who were MH, a considerably higher proportion had a low 10-year CVD risk, and lower proportions had a high CVD risk regardless of BMI levels: 79.3% (157/198) of participants who were MHO had a low risk, whereas a high risk was only observed among 5.1% (10/198) of participants who were MHO. A similar pattern was noted in the MHNW and MHOW groups.

Table 2. Levels of and distribution of the Framingham Risk Score (FRS) among each obesity phenotype.

Distribution	MHNW[a] (n=2548)	MHOW[b] (n=1135)	MHO[c] (n=198)	MUNW[d] (n=955)	MUOW[e] (n=981)	MUO[f] (n=459)
FRS distribution, mean (SD)	7.43 (8.39)	8.54 (9.42)	7.54 (7.91)	14.16 (13.01)	15.63 (14.08)	15.98 (14.42)
FRS categories, n (%)						
Low	1930 (75.75)	831 (73.22)	157 (79.3)	479 (50.2)	452 (46.1)	210 (45.8)
Moderate	412 (16.17)	196 (17.27)	31 (15.7)	247 (25.9)	270 (27.5)	123 (26.8)
High	206 (8.08)	108 (9.52)	10 (5.1)	229 (24)	259 (26.4)	126 (27.5)

[a]MHNW: metabolically healthy normal weight.

[b]MHOW: metabolically healthy overweight.

[c]MHO: metabolically healthy obese.

[d]MUNW: metabolically unhealthy normal weight.

[e]MUOW: metabolically unhealthy overweight.

[f]MUO: metabolically unhealthy obese.

BN Development

BN modeling was used to estimate the 10-year CVD risk among various obesity phenotypes. By using a whitelist and blacklist from prior expert knowledge, an averaged BN was constructed by learning 300 bootstrapped networks from the data and further retaining the arcs with an appearing frequency of at least 50%, as shown in Figure 1A. This BN model describes the interrelationships between demographic factors, behavioral factors, CVD risk factors, obesity phenotypes, and FRS

categories, as well as the relationship between risk factors and demographic covariates. All the directions of the arcs seem to be well-established, and this could be attributed to the layering approach (whitelist and blacklist), which implements certain restrictions on the arc directions.

Then, a further examination was performed using the arc strength criteria to simplify the complexity of the BN, with little loss of information in the process. The final BN was obtained with an arc strength threshold >0.85 (ie, arcs appear with a frequency of at least 0.85 among the 300 bootstrapped

networks), as depicted in Figure 1B (the direct comparison of arcs between the averaged BN and simplified BN is shown in Multimedia Appendix 1). All probability distributions are represented in the nodes, and the probabilistic dependencies are indicated by direct edges connecting the nodes. The connections between 10-year CVD risk, its risk factors, and obesity phenotypes were established by a complex network structure and assumed to be dependent, in which direct connections among metabolic health status, BMI level, age, smoking status, and CVD risk were identified (Figure 1B), together with an indirect link between sex and 10-year CVD risk through smoking status. In addition, metabolic health status and sex were directly connected with 7 and 5 covariates, respectively, indicating that they had the most children nodes, implying a sex-specific relationship. The interrelationships between various CVD risk factors are also presented in Figure 1B. For instance, eGFR was related to sex and age; hypertension and hyperuricemia were both associated with metabolic health status and sex; and TC was influenced by eGFR, dyslipidemia, and metabolic health status.

Figure 1. The directed acyclic graph (DAG) underlying the Bayesian network learned from 10-year cardiovascular disease (CVD) risk, the covariates, metabolic health, and obesity status. (A) Averaged DAG with strength of arcs >0.5; (B) simplified DAG derived from the averaged DAG after retaining arcs with a strength >0.85. eGFR: estimated glomerular filtration rate; hsCRP: high-sensitivity C-reactive protein; LDL-C: low-density lipoprotein cholesterol; TC: total cholesterol.
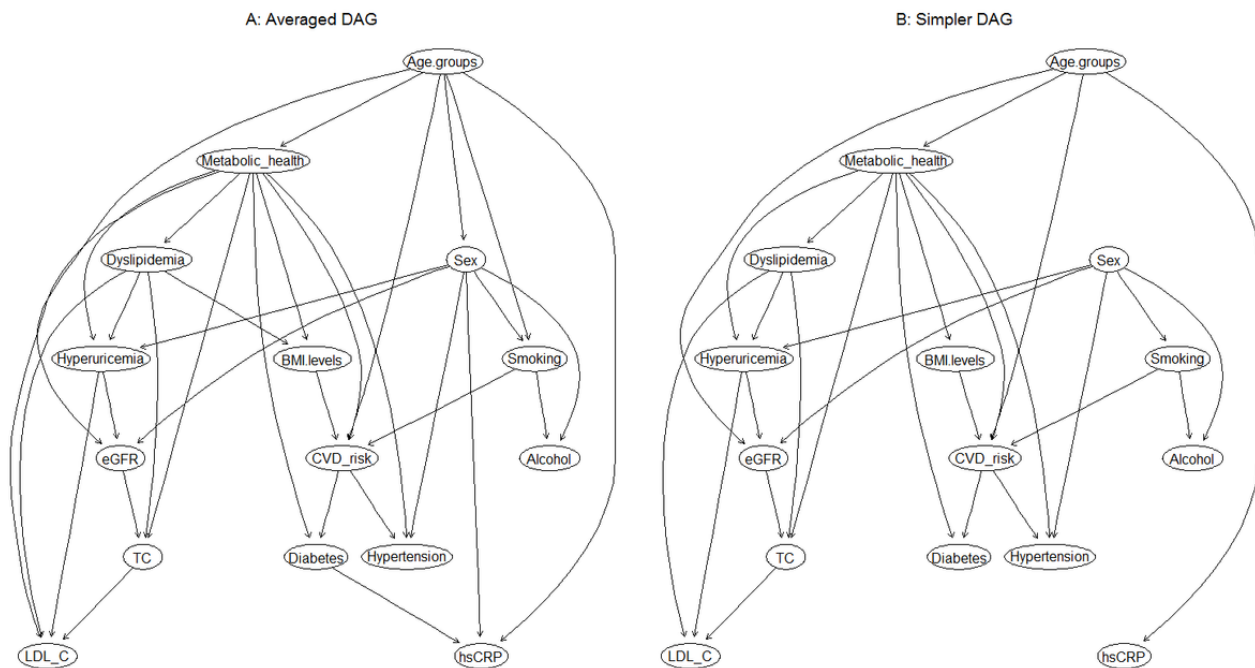
BN Reasoning

BN reasoning was performed to estimate the conditional probabilities of the 10-year CVD risk, given various evidence from the well-built BN model. Since age had a significant modifying effect on the probability distribution of CVD, the variation in the conditional probabilities was estimated through the different age groups (Figure 2). The conditional probability of high CVD risk increased as age progressed, with a greater rise from 30 to 40 years (0.43%, 95% CI 0.2-0.87) to 40 to 50 years (2.25%, 95% CI 1.75-2.88) and then to 50 to 60 years (16.13%, 95% CI 14.86-17.5). Furthermore, more than half of the participants aged ≥70 years had a high CVD risk (52.02%, 95% CI 47.62-56.38). A similar pattern was also observed for moderate CVD risk among the different age groups, with a steady increase in the conditional probability observed with increasing age. In contrast, the probability of low CVD risk decreased from 98% (95% CI 97.24 to 98.55) for the 30 to 40 years age group to 58.05% (95% CI 56.27 to 59.81) for the 50 to 60 years age group to only 14.92% (95% CI 12.04 to 18.34) for those aged ≥70 years.

The probability distributions of CVD risk were updated when providing evidence of BMI level and metabolic health status from the BN model (Figure 3). Among the participants with favorable metabolic health profiles, the conditional probability of having moderate or high CVD risk ranged from 15.3% and 7.01%, respectively, for the participants with normal weight (ie, MHNW) to 18.6% and 10.47%, respectively, for their obese counterparts (ie, MHO). In contrast, within the same BMI levels, the probabilities were 25.28% and 21.74%, respectively, for participants who were normal weight with a MU status (ie, MUNW) and further increased to 24.45% and 34.48%, respectively, among participants who were MUO. In addition, the conditional probabilities of low CVD risk exhibited a substantial decline from 77.69% to 52.98% when the metabolic health status of participants who were normal weight (ie, MHNW) became unfavorable (MUNW).

Subgroup analyses were conducted by providing further evidence of a sex factor in the BN model; these analyses are summarized in Figure 4. In men, the conditional probabilities of high CVD risk were nearly doubled among participants who were MH, regardless of BMI level, with probabilities ranging from 12.28% to 21.92% for men who were MHNW and MHO,

respectively. This indicates that men who were MH had twice the chance of high CVD risk when compared with their general population counterparts within the same obesity phenotype. Similarly, the men who were MU also had an increased CVD risk, where male participants who were MUO were nearly 2-fold more likely to have a high CVD risk than their MHO counterparts, with a conditional probability up to 48.21% (95% CI 42.92 to 53.55), whereas these probabilities were raised by a factor of ≥2.5 among men who were MUNW and MUOW when compared with their MH counterparts within the same BMI levels (MUNW vs MHNW: 32.18% vs 12.28%; MUOW vs MHOW: 43.17% vs 15.4%).

In contrast, female participants with favorable metabolic health profiles had substantially lower conditional probability estimates of moderate and high CVD risk, irrespective of their BMI level, with probabilities of only 10.61% and 2.02% in women who were MHO. Similarly, only approximately one-fifth of the women who were MUO (22.7%, 95% CI 18.83-27.11) and one-tenth of the women who were MU and nonobese (13% for women who were MUNW and 11.91% for women who were MUOW) had a high risk of developing CVD, whereas women who were MH (irrespective of obesity) were more than half as likely to have a low CVD risk, with corresponding conditional probabilities of 64.74% (95% CI 61.43 to 67.92), 58.84% (95% CI 55.5 to 62.11), and 54.08% (95% CI 49.13-58-95) for women who were MUNW, MUOW, and MUO, respectively.

Figure 2. Conditional probabilities (in percentage) and 95% CIs of low, moderate, and high 10-year cardiovascular disease (CVD) risk in different age groups in Chinese adults.
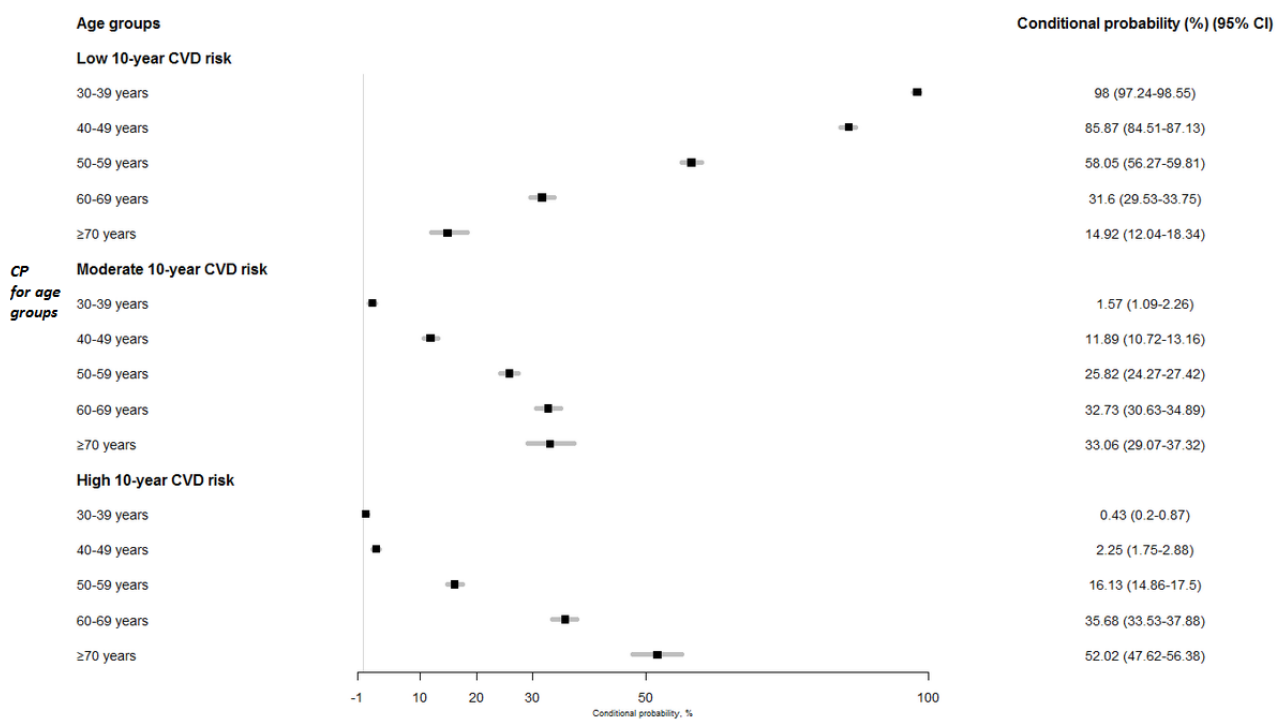
Figure 3. Conditional probabilities (in percentage) and 95% CIs of low, moderate, and high 10-year cardiovascular disease (CVD) risk in different obesity phenotypes. MHNW: metabolically healthy normal weight; MHO: metabolically healthy obese; MHOW: metabolically healthy overweight; MUNW: metabolically unhealthy normal weight; MUO: metabolically unhealthy obese; MUOW: metabolically unhealthy overweight.
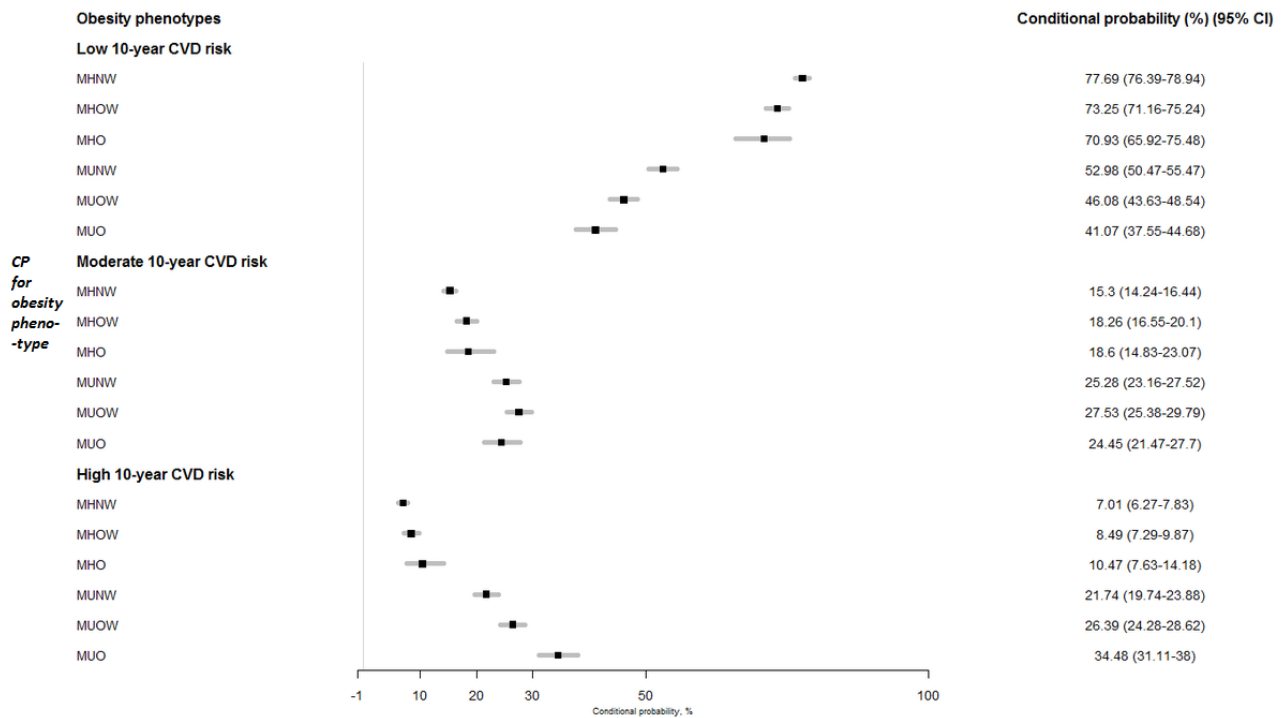
Figure 4. Conditional probabilities (in percentage) and 95% CIs of low, moderate, and high 10-year cardiovascular disease (CVD) risk in different obesity phenotypes by sex. MHNW: metabolically healthy normal weight; MHO: metabolically healthy obese; MHOW: metabolically healthy overweight; MUNW: metabolically unhealthy normal weight; MUO: metabolically unhealthy obese; MUOW: metabolically unhealthy overweight.
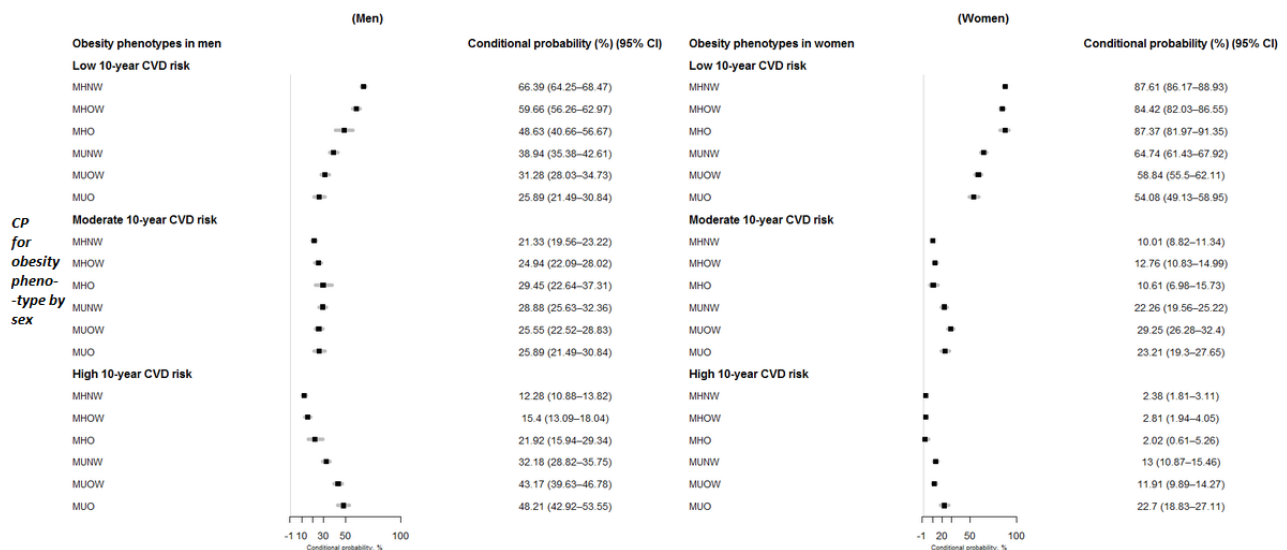
Discussion

Principal Findings

To the best of our knowledge, this is the first large-scale study that has applied the BN modeling approach to investigate the probabilistic relationship between different metabolic and obesity phenotypes and the 10-year CVD risk in Chinese adults. Individuals who were MHO had an increased probability (10.47%) of high CVD risk, and this probability doubled among participants who were MUNW and tripled for participants who

were MUO. Furthermore, an important gap in conditional probabilities was found between the two sexes within each obesity phenotype, suggesting a prominent modifying effect of sex on this relationship. The proposed DAG structure in the network represents a relevant step in understanding the complex interrelationships between the variables investigated and provides a self-descriptive and contextualized picture of these complex interrelationships in the Chinese population.

Compared with previous studies, this work offers a more comprehensive picture that simultaneously describes the

complex interrelationships between metabolic healthy or unhealthy phenotypes and 10-year CVD risk across BMI categories, as well as the relationships among CVD-related risk factors. Although several prospective cohort studies have focused on exploring the specific associations between CVD risk and metabolic health status and BMI categories [14,45-47], or the dynamic risk in the transition from one phenotype to another [48-50], this work aimed to provide a general framework to understand the multiple association processes that can emerge from the complex interrelationships between these factors. In fact, compared with standard studies, the greatest advantage and strength of this BN methodological approach is the graphical clarification and visualization of the most probable pathways in these relationships from a multi-dependent perspective, without affecting the interpretability of the network. Another advantage of BN modeling is that owing to the Markov blanket theory, complex models can be divided into a collection of simpler models that are mathematically tractable and computationally simpler. For instance, according to our BN reasoning model (Figure 1B), when a participant who was MH and normal weight became obese, the probability of developing a high 10-year CVD level increased from 7.01% to 10.47%, and this probability was tripled if the participant remained normal weight but had any ≥2 components of MetS. Furthermore, the probability increases up to 34.48% when the participant has both a MU status and is obese (ie, MUO; Figure 2). Moreover, remarkable heterogeneity in the associations between obesity phenotypes and CVD risk in relation to participant sex was observed. The conditional probability of having a high CVD risk was 2.02% and 22.7% among women who were MHO and MUO, respectively, whereas it rose to as high as 21.92% and 48.21% among men with the corresponding obesity phenotypes. Therefore, BN modeling enabled us to achieve an integrated view of CVD risk among the various obesity phenotypes in the context of other risk factors. It also allows for systematic reasoning in the diagnostic process with easy interpretability.

Comparison With Prior Work

Many studies have investigated the associations between obesity phenotypes and CVD outcomes, including myocardial infarction [51], HF [45,52], coronary heart disease [45,53], atrial fibrillation [54], and cerebrovascular disease [45]. This study confirms an elevated risk of CVD in people classified as MHOW or MHO when compared with their counterparts who are of normal weight and are MH. This is in line with previous studies in Asian [50,55], European [46], and American populations [56]. In the Danish prospective Inter99 study, Hansen et al [47] found that men who were MHO had a 3-fold increased risk of incident CVD compared with their MHNW counterparts, and a similar and significant augmentation of CVD risk was also observed in men who were MU, with a 2.2-fold increased risk in men who were MUNW and an approximately 3-fold increase in men who were MUOW and MUO, respectively, during the 10-year follow-up. In contrast, the increased risk among women who were MH was not significant, irrespective of BMI level, when compared with their MHNW counterparts. In fact, an inverted U-shaped relationship was observed between women who were MU and CVD risk across the BMI levels, and a

significant 2.3-fold increased risk was only observed among women who were MUOW compared with women who were MHNW [47]. The Nurses' Health Study, which included 90,257 American women aged 30 to 55 years without CVD or cancer history [56], indicated that women with either higher BMI levels or MU status were at a significantly increased risk compared with their MHNW counterparts after a follow-up of 30 years. In addition, women who were MH had a substantially lower risk of CVD than women with pre-existing metabolic conditions across all BMI groups. Consistently, the Whitehall 2 study also found that, compared with the MHNW phenotype, the risk of incident CVD was significantly elevated in the 5 other phenotypes during a median follow-up of 17.4 years, with adjusted hazard ratios (HRs) of 1.95 for MHO and 2.44 for MUO. Similarly, the participants who were MU had a higher risk than their MH counterparts, irrespective of BMI levels, by a factor of 2 for the nonobese BMI level and 1.2 for the obese BMI level [46]. Similarly, the findings from several Chinese prospective cohort studies, with either short- or long-term follow-ups [48,57,58], are consistent with those obtained in Western populations. The largest prospective cohort study, the China Kadoorie Biobank study, which comprised 458,246 Chinese participants without any history of CVD or cancer, found that after 10 years of follow-up, individuals who were baseline MHO had an 8% higher risk of developing CVD compared with their MHNW counterparts, and the risk for individuals who were MU was significantly higher across all BMI categories by a factor of 1.6 [48]. Of note, a very recent meta-analysis of 23 prospective cohort studies and 4,492,723 participants confirmed an elevated risk of CVD in individuals classified as MHOW or MHO when compared with their MHNW counterparts by a factor of 1.34 and 1.5, respectively. This increased risk remained statistically significant among individuals who were MHOW or MHO when defining metabolic health status with a strict definition (ie, having no metabolic risk factors) [14]. This indicates the potential nonexistence of the MHOW or MHO concept. Another important observation, as described in the meta-analyses by Kramer et al [59], Fan et al [60], Eckel et al [61], Zheng et al [62], Ortega et al [63], and others, is that the high risk of CVD associated with MHO appears to be sustained over a long-term follow-up (≥15 years). Several mechanisms could explain the potential association between MHO and the risk of CVD. Individuals who were MHO or MHOW may have higher odds of subclinical CVD and diabetes, which increases their likelihood of developing CVD in the future [64].

Smoking status and age have well-known causal effects on long-term CVD risk [65-67]. The current results suggest that smoking status and age are directly associated with 10-year CVD risk and seem to mediate the effect of sex and other variables included in our BN model. To the best of our knowledge, few studies have focused on sex-specific differences in the relationship between obesity phenotypes and CVD risk, although sex is a very important factor [68,69]. In this, men who were obese had an elevated probability of high 10-year CVD risk when compared with their female peers, irrespective of metabolic health status. Although there are some discrepancies in the way that obesity and metabolic health status have been defined, the current findings of sex-specific

differences are consistent with other large prospective studies in Chinese, European, and US samples [47,48,70]. For instance, Danish men who were MHO and MUO had a 3.1- and 2.7-fold increased CVD risk compared with their MHNW counterparts, whereas this increased risk was only by a factor of 1.8 in women within the same comparison of phenotypes [47]. Similarly, the China Kadoorie Biobank study found that men had a 1.09 times higher risk of CVD subtypes when compared with women within the MHO phenotype and a 1.3 times higher increased risk within the MUO phenotype [48]. This finding was also supported by a recent pooled analysis of prospective cohort studies, which revealed that men who were MHO had a 1.26 times increased risk compared with women who were MHO (HR: 2.15 vs 1.71) [14].

The present results are also in agreement with those of previous cohort studies that described the progression of CVD risk with aging. The China Kadoorie Biobank study demonstrated a clear age-specific pattern in CVD risk, irrespective of obesity phenotypes [48]. A steady rise in CVD risk was noted from age 30 to 49 years to 50 to 59 years, and this risk was substantially increased for individuals aged ≥60 years within each obesity phenotype. Individuals who were MUO aged 70 to 79 years had the highest risk of developing CVD events among all obesity phenotypes, with a 13.86-fold higher risk when taking individuals with MHNW at age 30 to 49 years as the reference group. Similarly, this risk was higher among participants who were MHNW aged ≥70 years compared with their counterparts aged 30 to 49 years. The current BN model applied to a population-based Chinese cohort showed a concave-shaped progression in the conditional probabilities for high CVD risk through the different age intervals together with a stronger increasing rate in the conditional probability after the age of 50 years (Figure 2). Clearly, the use of such BN modeling with respect to prior knowledge could provide quantitative descriptions of direct links between CVD and its related risk factors by intuitive reasoning. More importantly, such modeling is suited to exploring indirect links through mediators and testing novel hypotheses by simulation.

The CVD risk in individuals who are MU with normal weight remains underinvestigated. A large pan-European prospective study of 8 European countries found that after a median follow-up of 12.2 years [53], the presence of metabolic abnormalities was associated with an increased risk of CHD at all levels of adiposity; more precisely, the MUNW phenotype had twice the risk of CHD compared with their MH counterparts. This finding is supported by recent data from the Women's Health Initiative Study [70] and a Korean prospective study [54]. Interestingly, several studies have demonstrated that the CVD risk in individuals who are MU is markedly higher than that of their MH counterparts across all BMI categories [53]. Similarly, when using the MHO group as a reference, the MU nonobese group was found to be at increased risk of atrial fibrillation, although this difference was not statistically significant [54]. The current results are in agreement with previous studies and contribute to the evidence indicating that individuals who are MUNW are at considerably higher CVD risk compared with their peers who are MHO, regardless of sex, and it seems reasonable to suggest that individuals who are

overweight or obese without metabolic abnormalities are at intermediate CVD risk, at a level between individuals who are healthy normal weight and individuals who are MU [53].

Another potential explanation for these findings may be related to the transient status of MHO during long-term follow-up. Indeed, several recent studies have considered and identified CVD risk according to the concurrent transition of metabolic health and weight status during different follow-up periods [49,50,56]. In one study, the recovery of MH status among individuals who were baseline MUNO was significantly associated with a decreased risk of CVD outcomes, whereas the transition from an MH status to an unhealthy status among the individuals who were baseline MUNO was related to adverse CVD outcomes [50]. These findings were also confirmed by Bae et al [49] using a nationally representative cohort study of 205,394 middle-aged Korean men and women who were followed up for 6 years. Interestingly, among the initial participants who were MHO, those who became MUNO had a 1.41-fold higher CVD risk compared with those who remained MHO. Together, this evidence strongly suggests a greater role of MetS than obesity in CVD risk, with longer exposure to a MU status leading to a much higher vascular risk [48]. Moreover, another nationwide population-based cohort study revealed that transition to the MUNO phenotype was associated with an 80% higher HR for HF among individuals who were MHO at baseline during a short-term follow-up (3.7 years). Conversely, individuals who transitioned from MHO to MH nonobese had a lower HR for HF than those who remained in the MHO category [52]. This could imply that restoration from an obese state to a nonobese state while maintaining metabolic health may have a protective effect against incident HF. However, as emphasized by Gao et al [48] in the largest Asian cohort study to date on the transitions between various obesity phenotypes over a longer follow-up, long-term maintenance of metabolic health is difficult for individuals of any BMI level, including individuals who are overweight and normal weight; therefore, more attention should be paid to maintaining metabolic health regardless of body weight. In addition, there should be a clinical focus on the treatment of metabolic disorders for CVD risk prevention. Similarly, efforts should be made to prevent the conversion of MHO to MUO and the development of MetS and subsequent CVD caused by obesity [71,72].

Limitations

There are several limitations to this study and the newly identified networks that should be noted. First, despite the longitudinal design of the CHNS survey, this study analyzed observational and cross-sectional data; the directions between nodes or variables only represent probability dependencies, not causal relationships. Further cohort studies combined with various aspects of professional knowledge are warranted to establish and clarify causality [73]. In addition, the sample only comprised Chinese participants; thus, the generalizability of these results to a wider population should be undertaken with caution. In addition, physical activity and fitness were lacking in the proposed BN modeling, although these 2 factors may be more important than weight in assessing CVD risk, as emphasized by Lavie et al [74]. In fact, the information on physical activity and fitness was not exhaustively provided from

the CHNS data; thus, they were not included in this study for minimizing the potential bias but will be well-considered in further studies.

Nevertheless, the notable strengths and contributions of this study should be mentioned. For example, the population-based design, large sample size, and rigorous data collection quality guarantee reasonable statistical power and robust probabilistic relationships. Second, the BN modeling approach offers compelling application prospects in general medicine. BN is not only useful for handling a large number of variables with or without prior knowledge of the interactions or interdependencies between them [38] but also provides an appealing visual presentation and quantitative reasoning that can be used to explore the interrelationships among these factors and test novel hypotheses.

Conclusions

The BN modeling approach was applied to investigate the relationships between different CVD risk factors and metabolic health and obesity status using Chinese population-based survey data. Network modeling is useful for integrating expert knowledge and observational data, allowing easy identification of probabilistic dependencies and conditional independencies between variables through graphical representation. This study provides evidence that increased CVD risk progresses depending on the varying magnitude of metabolic abnormalities and BMI. Furthermore, several potential modifying factors were identified, including sex, that may affect previous probabilistic interrelationships. Owing to the multifactorial nature of CVD, these empirical findings using the BN approach are of special interest, both from a theoretical and practical point of view, and may help in refining appropriate target populations and relevant risk factors for managing future CVD risk.

Acknowledgments

Authors' Contributions

ST and YL were involved in the conceptualization of the study. Data acquisition was performed by MB, YB, and XC. Statistical analysis was conducted by ST, YB, and XC. Investigation was conducted by MB and YL. ST and MB wrote the original draft. YL was involved in reviewing and editing the paper. All authors contributed to the manuscript and approved the submitted version.

Conflicts of Interest

None declared.

Multimedia Appendix 1
The directed acyclic (DAG) graph underlying the Bayesian network learned from 10-year cardiovascular disease risk, covariates, metabolic health, and obesity status. (A) Averaged DAG with arcs >0.5; (B) simplified DAG derived from the averaged DAG after retaining arcs with a strength >0.85.
[PNG File , 153 KB - medinform_v10i3e33026_app1.png]

References

1. Roth GA, Mensah GA, Johnson CO, Addolorato G, Ammirati E, Baddour LM, GBD-NHLBI-JACC Global Burden of Cardiovascular Diseases Writing Group. Global burden of cardiovascular diseases and risk factors, 1990-2019: update from the GBD 2019 study. J Am Coll Cardiol 2020 Dec 22;76(25):2982-3021 [FREE Full text] [doi: 10.1016/j.jacc.2020.11.010] [Medline: 33309175]

2. Zhao D, Liu J, Wang M, Zhang X, Zhou M. Epidemiology of cardiovascular disease in China: current features and implications. Nat Rev Cardiol 2019 Apr;16(4):203-212. [doi: 10.1038/s41569-018-0119-4] [Medline: 30467329]

3. Liu S, Li Y, Zeng X, Wang H, Yin P, Wang L, et al. Burden of cardiovascular diseases in China, 1990-2016: findings from the 2016 global burden of disease study. JAMA Cardiol 2019 Apr 01;4(4):342-352 [FREE Full text] [doi: 10.1001/jamacardio.2019.0295] [Medline: 30865215]

4. Yatsuya H, Li Y, Hilawe EH, Ota A, Wang C, Chiang C, et al. Global trend in overweight and obesity and its association with cardiovascular disease incidence. Circ J 2014;78(12):2807-2818 [FREE Full text] [doi: 10.1253/circj.cj-14-0850] [Medline: 25391910]

5. O'Brien EC, Fosbol EL, Peng SA, Alexander KP, Roe MT, Peterson ED. Association of body mass index and long-term outcomes in older patients with non-ST-segment-elevation myocardial infarction: results from the CRUSADE Registry. Circ Cardiovasc Qual Outcomes 2014 Jan;7(1):102-109. [doi: 10.1161/CIRCOUTCOMES.113.000421] [Medline: 24326936]

6. Lahey R, Khan SS. Trends in obesity and risk of cardiovascular disease. Curr Epidemiol Rep 2018 Sep;5(3):243-251 [FREE Full text] [doi: 10.1007/s40471-018-0160-1] [Medline: 30705802]

7.   Chen Z, Iona A, Parish S, Chen Y, Guo Y, Bragg F, China Kadoorie Biobank collaborative group. Adiposity and risk of ischaemic and haemorrhagic stroke in 0·5 million Chinese men and women: a prospective cohort study. Lancet Glob Health 2018 Jun;6(6):e630-e640 [FREE Full text] [doi: 10.1016/S2214-109X(18)30216-X] [Medline: 29773119]

8.   Alberti KG, Eckel RH, Grundy SM, Zimmet PZ, Cleeman JI, Donato KA, International Diabetes Federation Task Force on EpidemiologyPrevention, Hational Heart, Lung,Blood Institute, American Heart Association, World Heart Federation, International Atherosclerosis Society, International Association for the Study of Obesity. Harmonizing the metabolic syndrome: a joint interim statement of the International Diabetes Federation Task Force on Epidemiology and Prevention; National Heart, Lung, and Blood Institute; American Heart Association; World Heart Federation; International Atherosclerosis Society; and International Association for the Study of Obesity. Circulation 2009 Oct 20;120(16):1640-1645. [doi: 10.1161/CIRCULATIONAHA.109.192644] [Medline: 19805654]

9.   Stefan N, Häring HU, Hu FB, Schulze MB. Metabolically healthy obesity: epidemiology, mechanisms, and clinical implications. Lancet Diabetes Endocrinol 2013 Oct;1(2):152-162. [doi: 10.1016/S2213-8587(13)70062-7] [Medline: 24622321]

10. Yang HK, Han K, Kwon H, Park Y, Cho J, Yoon K, et al. Obesity, metabolic health, and mortality in adults: a nationwide population-based study in Korea. Sci Rep 2016 Jul 22;6:30329 [FREE Full text] [doi: 10.1038/srep30329] [Medline: 27445194]

11. Task Force Members, Montalescot G, Sechtem U, Achenbach S, Andreotti F, Arden C, ESC Committee for Practice Guidelines, Document Reviewers, et al. 2013 ESC guidelines on the management of stable coronary artery disease: the Task Force on the management of stable coronary artery disease of the European Society of Cardiology. Eur Heart J 2013 Oct;34(38):2949-3003. [doi: 10.1093/eurheartj/eht296] [Medline: 23996286]

12. Writing Group Members, Mozaffarian D, Benjamin EJ, Go AS, Arnett DK, Blaha MJ, American Heart Association Statistics Committee, Stroke Statistics Subcommittee. Executive summary: heart disease and stroke statistics--2016 update: a report from the American Heart Association. Circulation 2016 Jan 26;133(4):447-454. [doi: 10.1161/CIR.0000000000000366] [Medline: 26811276]

13. Badawi A, Di Giuseppe G, Arora P. Cardiovascular disease risk in patients with hepatitis C infection: results from two general population health surveys in Canada and the United States (2007-2017). PLoS One 2018;13(12):e0208839 [FREE Full text] [doi: 10.1371/journal.pone.0208839] [Medline: 30540839]

14. Opio J, Croker E, Odongo GS, Attia J, Wynne K, McEvoy M. Metabolically healthy overweight/obesity are associated with increased risk of cardiovascular disease in adults, even in the absence of metabolic risk factors: a systematic review and meta-analysis of prospective cohort studies. Obes Rev 2020 Dec;21(12):e13127. [doi: 10.1111/obr.13127] [Medline: 32869512]

15. Scutari M, Vitolo C, Tucker A. Learning Bayesian networks from big data with greedy search: computational complexity and efficient implementation. Stat Comput 2019 Feb 15;29(5):1095-1108. [doi: 10.1007/s11222-019-09857-1]

16. Jensen F, Nielsen T. Bayesian Networks and Decision Graphs. Cham: Springer; 2001.

17. Gupta A, Slater JJ, Boyne D, Mitsakakis N, Béliveau A, Druzdzel MJ, et al. Probabilistic graphical modeling for estimating risk of coronary artery disease: applications of a flexible machine-learning method. Med Decis Making 2019 Nov;39(8):1032-1044. [doi: 10.1177/0272989X19879095] [Medline: 31619130]

18. Fuster-Parra P, Tauler P, Bennasar-Veny M, Ligęza A, López-González AA, Aguiló A. Bayesian network modeling: a case study of an epidemiologic system analysis of cardiovascular risk. Comput Methods Programs Biomed 2016 Apr;126:128-142. [doi: 10.1016/j.cmpb.2015.12.010] [Medline: 26777431]

19. Badawi A, Di Giuseppe G, Gupta A, Poirier A, Arora P. Bayesian network modelling study to identify factors influencing the risk of cardiovascular disease in Canadian adults with hepatitis C virus infection. BMJ Open 2020 May 05;10(5):e035867 [FREE Full text] [doi: 10.1136/bmjopen-2019-035867] [Medline: 32371519]

20. Yan S, Li J, Li S, Zhang B, Du S, Gordon-Larsen P, et al. The expanding burden of cardiometabolic risk in China: the China Health and Nutrition Survey. Obes Rev 2012 Sep;13(9):810-821 [FREE Full text] [doi: 10.1111/j.1467-789X.2012.01016.x] [Medline: 22738663]

21. Tian S, Liu Y, Feng A, Zhang S. Sex-specific differences in the association of metabolically healthy obesity with hyperuricemia and a network perspective in analyzing factors related to hyperuricemia. Front Endocrinol (Lausanne) 2020;11:573452 [FREE Full text] [doi: 10.3389/fendo.2020.573452] [Medline: 33123092]

22. Ye X, Liu X, Song D, Zhang X, Zhu B, Wei L, et al. Estimating glomerular filtration rate by serum creatinine or/and cystatin C equations: an analysis of multi-centre Chinese subjects. Nephrology (Carlton) 2016 May;21(5):372-378. [doi: 10.1111/nep.12636] [Medline: 26427030]

23. James PA, Oparil S, Carter BL, Cushman WC, Dennison-Himmelfarb C, Handler J, et al. 2014 evidence-based guideline for the management of high blood pressure in adults: report from the panel members appointed to the Eighth Joint National Committee (JNC 8). JAMA 2014 Feb 05;311(5):507-520. [doi: 10.1001/jama.2013.284427] [Medline: 24352797]

24. Liu L, Writing Group of 2010 Chinese Guidelines for the Management of Hypertension. [2010 Chinese guidelines for the management of hypertension]. Zhonghua Xin Xue Guan Bing Za Zhi 2011 Jul;39(7):579-615. [Medline: 22088239]

25. National Cholesterol Education Program (NCEP) Expert Panel on Detection, Evaluation,Treatment of High Blood Cholesterol in Adults (Adult Treatment Panel III). Third report of the national cholesterol education program (NCEP) expert panel on

detection, evaluation, and treatment of high blood cholesterol in adults (adult treatment panel III) final report. Circulation 2002 Dec 17;106(25):3143-3421. [Medline: 12485966]

26. Zhang W, Doherty M, Bardin T, Pascual E, Barskova V, Conaghan P, EULAR Standing Committee for International Clinical Studies Including Therapeutics. EULAR evidence based recommendations for gout. Part II: management. Report of a task force of the EULAR Standing Committee for International Clinical Studies Including Therapeutics (ESCISIT). Ann Rheum Dis 2006 Oct;65(10):1312-1324 [FREE Full text] [doi: 10.1136/ard.2006.055269] [Medline: 16707532]

27. D'Agostino RB, Vasan RS, Pencina MJ, Wolf PA, Cobain M, Massaro JM, et al. General cardiovascular risk profile for use in primary care: the Framingham Heart Study. Circulation 2008 Feb 12;117(6):743-753. [doi: 10.1161/CIRCULATIONAHA.107.699579] [Medline: 18212285]

28. Bosomworth NJ. Practical use of the Framingham risk score in primary prevention: Canadian perspective. Can Fam Physician 2011 Apr;57(4):417-423 [FREE Full text] [Medline: 21626897]

29. Anderson TJ, Grégoire J, Pearson GJ, Barry AR, Couture P, Dawes M, et al. 2016 canadian cardiovascular society guidelines for the management of dyslipidemia for the prevention of cardiovascular disease in the adult. Can J Cardiol 2016 Nov;32(11):1263-1282. [doi: 10.1016/j.cjca.2016.07.510] [Medline: 27712954]

30. Novo S, Carità P, Lo Voi A, Muratori I, Tantillo R, Corrado E, et al. Impact of preclinical carotid atherosclerosis on global cardiovascular risk stratification and events in a 10-year follow-up: comparison between the algorithms of the Framingham Heart Study, the European SCORE and the Italian 'Progetto Cuore'. J Cardiovasc Med (Hagerstown) 2019 Feb;20(2):91-96. [doi: 10.2459/JCM.0000000000000740] [Medline: 30557211]

31. Jamthikar A, Gupta D, Cuadrado-Godia E, Puvvula A, Khanna NN, Saba L, et al. Ultrasound-based stroke/cardiovascular risk stratification using Framingham Risk Score and ASCVD Risk Score based on "Integrated Vascular Age" instead of "Chronological Age": a multi-ethnic study of Asian Indian, Caucasian, and Japanese cohorts. Cardiovasc Diagn Ther 2020 Aug;10(4):939-954 [FREE Full text] [doi: 10.21037/cdt.2020.01.16] [Medline: 32968652]

32. Zhou J, Gao Q, Wang J, Zhang M, Ma J, Wang C, et al. Comparison of coronary heart disease risk assessments among individuals with metabolic syndrome using three diagnostic definitions: a cross-sectional study from China. BMJ Open 2018 Oct 25;8(10):e022974 [FREE Full text] [doi: 10.1136/bmjopen-2018-022974] [Medline: 30366915]

33. Shillinglaw B, Viera AJ, Edwards T, Simpson R, Sheridan SL. Use of global coronary heart disease risk assessment in practice: a cross-sectional survey of a sample of U.S. physicians. BMC Health Serv Res 2012 Jan 24;12:20 [FREE Full text] [doi: 10.1186/1472-6963-12-20] [Medline: 22273080]

34. Chen C, Lu FC, Department of Disease Control Ministry of Health, PR China. The guidelines for prevention and control of overweight and obesity in Chinese adults. Biomed Environ Sci 2004;17 Suppl:1-36. [Medline: 15807475]

35. The Asia-pacific Perspective : Redefining Obesity and Its Treatment. Sydney, New South Wales: Health Communications Australia; 2000.

36. Expert Panel on Detection, Evaluation,Treatment of High Blood Cholesterol in Adults. Executive Summary of The Third Report of The National Cholesterol Education Program (NCEP) expert panel on detection, evaluation, and treatment of high blood cholesterol in adults (Adult Treatment Panel III). JAMA 2001 May 16;285(19):2486-2497. [doi: 10.1001/jama.285.19.2486] [Medline: 11368702]

37. Constantinou AC, Fenton N, Marsh W, Radlinski L. From complex questionnaire and interviewing data to intelligent Bayesian network models for medical decision support. Artif Intell Med 2016 Feb;67:75-93 [FREE Full text] [doi: 10.1016/j.artmed.2016.01.002] [Medline: 26830286]

38. Chao Y, Wu H, Scutari M, Chen T, Wu C, Durand M, et al. A network perspective on patient experiences and health status: the Medical Expenditure Panel Survey 2004 to 2011. BMC Health Serv Res 2017 Aug 22;17(1):579 [FREE Full text] [doi: 10.1186/s12913-017-2496-5] [Medline: 28830413]

39. Amirkhani H, Rahmati M, Lucas PJ, Hommersom A. Exploiting experts' knowledge for structure learning of Bayesian networks. IEEE Trans Pattern Anal Mach Intell 2017 Nov;39(11):2154-2170. [doi: 10.1109/TPAMI.2016.2636828] [Medline: 28114005]

40. Sambo F, Di Camillo B, Franzin A, Facchinetti A, Hakaste L, Kravic J, et al. A Bayesian Network analysis of the probabilistic relations between risk factors in the predisposition to type 2 diabetes. Annu Int Conf IEEE Eng Med Biol Soc 2015;2015:2119-2122. [doi: 10.1109/EMBC.2015.7318807] [Medline: 26736707]

41. Glover F. Artificial intelligence, heuristic frameworks and tabu search. Manage Decision Econ 1990;11(5):365-375. [doi: 10.1002/mde.4090110512]

42. Schwarz G. Estimating the dimension of a model. Ann Stat 1978 Mar 1;6(2):461-464. [doi: 10.1214/aos/1176344136]

43. Scutari M. Learning Bayesian networks with the package. J Stat Softw 2010;35(3). [doi: 10.18637/jss.v035.i03]

44. R: A Language and Environment for Statistical Computing. Vienna, Austria: R Foundation for Statistical Computing; 2020.

45. Caleyachetty R, Thomas GN, Toulis KA, Mohammed N, Gokhale KM, Balachandran K, et al. Metabolically healthy obese and incident cardiovascular disease events among 3.5 million men and women. J Am Coll Cardiol 2017 Sep 19;70(12):1429-1437 [FREE Full text] [doi: 10.1016/j.jacc.2017.07.763] [Medline: 28911506]

46. Hinnouho G, Czernichow S, Dugravot A, Nabi H, Brunner EJ, Kivimaki M, et al. Metabolically healthy obesity and the risk of cardiovascular disease and type 2 diabetes: the Whitehall II cohort study. Eur Heart J 2015 Mar 01;36(9):551-559 [FREE Full text] [doi: 10.1093/eurheartj/ehu123] [Medline: 24670711]

47. Hansen L, Netterstrøm MK, Johansen NB, Rønn PF, Vistisen D, Husemoen LL, et al. Metabolically healthy obesity and ischemic heart disease: a 10-year follow-up of the Inter99 study. J Clin Endocrinol Metab 2017 Jun 01;102(6):1934-1942. [doi: 10.1210/jc.2016-3346] [Medline: 28323999]

48. Gao M, Lv J, Yu C, Guo Y, Bian Z, Yang R, China Kadoorie Biobank (CKB) Collaborative Group. Metabolically healthy obesity, transition to unhealthy metabolic status, and vascular disease in Chinese adults: a cohort study. PLoS Med 2020 Oct;17(10):e1003351 [FREE Full text] [doi: 10.1371/journal.pmed.1003351] [Medline: 33125374]

49. Bae YS, Choi S, Lee K, Son JS, Lee H, Cho MH, et al. Association of concurrent changes in metabolic health and weight on cardiovascular disease risk: a nationally representative cohort study. J Am Heart Assoc 2019 Sep 03;8(17):e011825 [FREE Full text] [doi: 10.1161/JAHA.118.011825] [Medline: 31451053]

50. Cho YK, Kang YM, Yoo JH, Lee J, Park J, Lee WJ, et al. Implications of the dynamic nature of metabolic health status and obesity on risk of incident cardiovascular events and mortality: a nationwide population-based cohort study. Metabolism 2019 Aug;97:50-56. [doi: 10.1016/j.metabol.2019.05.002] [Medline: 31071310]

51. Mirzababaei A, Djafarian K, Mozafari H, Shab-Bidar S. The long-term prognosis of heart diseases for different metabolic phenotypes: a systematic review and meta-analysis of prospective cohort studies. Endocrine 2019 Mar;63(3):439-462. [doi: 10.1007/s12020-019-01840-0] [Medline: 30671787]

52. Lee Y, Kim DH, Kim SM, Kim NH, Choi KM, Baik SH, et al. Hospitalization for heart failure incidence according to the transition in metabolic health and obesity status: a nationwide population-based study. Cardiovasc Diabetol 2020 Jun 13;19(1):77 [FREE Full text] [doi: 10.1186/s12933-020-01051-2] [Medline: 32534576]

53. Lassale C, Tzoulaki I, Moons KG, Sweeting M, Boer J, Johnson L, et al. Separate and combined associations of obesity and metabolic health with coronary heart disease: a pan-European case-cohort analysis. Eur Heart J 2018 Feb 01;39(5):397-406 [FREE Full text] [doi: 10.1093/eurheartj/ehx448] [Medline: 29020414]

54. Lee H, Choi E, Lee S, Han K, Rhee T, Park C, et al. Atrial fibrillation risk in metabolically healthy obesity: a nationwide population-based study. Int J Cardiol 2017 Aug 01;240:221-227. [doi: 10.1016/j.ijcard.2017.03.103] [Medline: 28385358]

55. Huang M, Wang M, Lin Y, Lin C, Lo K, Chang I, et al. The association between metabolically healthy obesity, cardiovascular disease, and all-cause mortality risk in Asia: a systematic review and meta-analysis. Int J Environ Res Public Health 2020 Feb 19;17(4):1320 [FREE Full text] [doi: 10.3390/ijerph17041320] [Medline: 32092849]

56. Eckel N, Li Y, Kuxhaus O, Stefan N, Hu FB, Schulze MB. Transition from metabolic healthy to unhealthy phenotypes and association with cardiovascular disease risk across BMI categories in 90 257 women (the Nurses' Health Study): 30 year follow-up from a prospective cohort study. Lancet Diabetes Endocrinol 2018 Sep;6(9):714-724. [doi: 10.1016/s2213-8587(18)30137-2]

57. Li L, Chen K, Wang A, Gao J, Zhao K, Wang H, et al. Cardiovascular disease outcomes in metabolically healthy obesity in communities of Beijing cohort study. Int J Clin Pract 2018 Sep 30:e13279. [doi: 10.1111/ijcp.13279] [Medline: 30269402]

58. Li H, He D, Zheng D, Amsalu E, Wang A, Tao L, et al. Metabolically healthy obese phenotype and risk of cardiovascular disease: results from the China Health and Retirement Longitudinal Study. Arch Gerontol Geriatr 2019;82:1-7. [doi: 10.1016/j.archger.2019.01.004] [Medline: 30710843]

59. Kramer CK, Zinman B, Retnakaran R. Are metabolically healthy overweight and obesity benign conditions?: a systematic review and meta-analysis. Ann Intern Med 2013 Dec 03;159(11):758-769. [doi: 10.7326/0003-4819-159-11-201312030-00008] [Medline: 24297192]

60. Fan J, Song Y, Chen Y, Hui R, Zhang W. Combined effect of obesity and cardio-metabolic abnormality on the risk of cardiovascular disease: a meta-analysis of prospective cohort studies. Int J Cardiol 2013 Oct 12;168(5):4761-4768. [doi: 10.1016/j.ijcard.2013.07.230] [Medline: 23972953]

61. Eckel N, Meidtner K, Kalle-Uhlmann T, Stefan N, Schulze MB. Metabolically healthy obesity and cardiovascular events: a systematic review and meta-analysis. Eur J Prev Cardiol 2016 Jun;23(9):956-966. [doi: 10.1177/2047487315623884] [Medline: 26701871]

62. Zheng R, Zhou D, Zhu Y. The long-term prognosis of cardiovascular disease and all-cause mortality for metabolically healthy obesity: a systematic review and meta-analysis. J Epidemiol Community Health 2016 Oct;70(10):1024-1031. [doi: 10.1136/jech-2015-206948] [Medline: 27126492]

63. Ortega FB, Cadenas-Sanchez C, Migueles JH, Labayen I, Ruiz JR, Sui X, et al. Role of physical activity and fitness in the characterization and prognosis of the metabolically healthy obesity phenotype: a systematic review and meta-analysis. Prog Cardiovasc Dis 2018;61(2):190-205. [doi: 10.1016/j.pcad.2018.07.008] [Medline: 30122522]

64. Dwivedi AK, Dubey P, Cistola DP, Reddy SY. Association between obesity and cardiovascular outcomes: updated evidence from meta-analysis studies. Curr Cardiol Rep 2020 Mar 12;22(4):25. [doi: 10.1007/s11886-020-1273-y] [Medline: 32166448]

65. Banks E, Joshy G, Korda RJ, Stavreski B, Soga K, Egger S, et al. Tobacco smoking and risk of 36 cardiovascular disease subtypes: fatal and non-fatal outcomes in a large prospective Australian study. BMC Med 2019 Jul 03;17(1):128 [FREE Full text] [doi: 10.1186/s12916-019-1351-4] [Medline: 31266500]

66. Duncan MS, Freiberg MS, Greevy RA, Kundu S, Vasan RS, Tindle HA. Association of smoking cessation with subsequent risk of cardiovascular disease. JAMA 2019 Aug 20;322(7):642-650 [FREE Full text] [doi: 10.1001/jama.2019.10298] [Medline: 31429895]

67. Costantino S, Paneni F, Cosentino F. Ageing, metabolism and cardiovascular disease. J Physiol 2016 Apr 15;594(8):2061-2073 [FREE Full text] [doi: 10.1113/JP270538] [Medline: 26391109]

68. Ventura-Clapier R, Dworatzek E, Seeland U, Kararigas G, Arnal J, Brunelleschi S, et al. Sex in basic research: concepts in the cardiovascular field. Cardiovasc Res 2017 Jun 01;113(7):711-724. [doi: 10.1093/cvr/cvx066] [Medline: 28472454]

69. Appelman Y, van Rijn BB, Ten Haaf ME, Boersma E, Peters SA. Sex differences in cardiovascular risk factors and disease prevention. Atherosclerosis 2015 Jul;241(1):211-218. [doi: 10.1016/j.atherosclerosis.2015.01.027] [Medline: 25670232]

70. Chen G, Arthur R, Iyengar NM, Kamensky V, Xue X, Wassertheil-Smoller S, et al. Association between regional body fat and cardiovascular disease risk among postmenopausal women with normal body mass index. Eur Heart J 2019 Sep 07;40(34):2849-2855 [FREE Full text] [doi: 10.1093/eurheartj/ehz391] [Medline: 31256194]

71. Lavie CJ, Deedwania P, Ortega FB. Obesity is rarely healthy. Lancet Diabetes Endocrinol 2018 Sep;6(9):678-679. [doi: 10.1016/S2213-8587(18)30143-8] [Medline: 29859910]

72. Deedwania P, Lavie CJ. Dangers and long-term outcomes in metabolically healthy obesity: the impact of the missing fitness component. J Am Coll Cardiol 2018 May 01;71(17):1866-1868 [FREE Full text] [doi: 10.1016/j.jacc.2018.02.057] [Medline: 29699612]

73. Nagarajan R, Scutari M, Lèbre S. Bayesian Networks in R. Cham: Springer; 2013.

74. Lavie CJ, Ozemek C, Carbone S, Katzmarzyk PT, Blair SN. Sedentary behavior, exercise, and cardiovascular health. Circ Res 2019 Mar;124(5):799-815. [doi: 10.1161/CIRCRESAHA.118.312669] [Medline: 30817262]

Abbreviations

BN: Bayesian network
BP: blood pressure
CHNS: China Health and Nutrition Survey
CVD: cardiovascular disease
DAG: directed acyclic graph
DBP: diastolic blood pressure
eGFR: estimated glomerular filtration rate
FPG: fasting plasma glucose
FRS: Framingham Risk Score
HDL-C: high-density lipoprotein cholesterol
HF: heart failure
HOMA-IR: homeostasis model assessment of insulin resistance
HR: hazard ratio
MetS: metabolic syndrome
MH: metabolically healthy
MHNW: metabolically healthy normal weight
MHO: metabolically healthy obese
MHOW: metabolically healthy overweight
MU: metabolically unhealthy
MUNW: metabolically unhealthy normal weight
MUO: metabolically unhealthy obese
MUOW: metabolically unhealthy overweight
SBP: systolic blood pressure
TC: total cholesterol
TG: triglyceride
UA: uric acid

<u>Original Paper</u>

A Data-Driven Algorithm to Recommend Initial Clinical Workup for Outpatient Specialty Referral: Algorithm Development and Validation Using Electronic Health Record Data and Expert Surveys

Wui Ip[1], MD; Priya Prahalad[1], MD, PhD; Jonathan Palma[2], MSc, MD; Jonathan H Chen[3,4], MD, PhD

[1]Department of Pediatrics, Stanford University School of Medicine, Palo Alto, CA, United States

[2]Neonatology & Perinatal Medicine, Orlando Health Winnie Palmer Hospital for Women & Babies, Orlando, FL, United States

[3]Department of Medicine, Stanford University School of Medicine, Palo Alto, CA, United States

[4]Stanford Center for Biomedical Informatics Research, Stanford, CA, United States

Corresponding Author:
Wui Ip, MD
Department of Pediatrics
Stanford University School of Medicine
453 Quarry Road, MC 5660
Palo Alto, CA, 94304
United States
Phone: 1 6507234000
Email: <u>wui@stanford.edu</u>

Abstract

Background: Millions of people have limited access to specialty care. The problem is exacerbated by ineffective specialty visits due to incomplete prereferral workup, leading to delays in diagnosis and treatment. Existing processes to guide prereferral diagnostic workup are labor-intensive (ie, building a consensus guideline between primary care doctors and specialists) and require the availability of the specialists (ie, electronic consultation).

Objective: Using pediatric endocrinology as an example, we develop a recommender algorithm to anticipate patients' initial workup needs at the time of specialty referral and compare it to a reference benchmark using the most common workup orders. We also evaluate the clinical appropriateness of the algorithm recommendations.

Methods: Electronic health record data were extracted from 3424 pediatric patients with new outpatient endocrinology referrals at an academic institution from 2015 to 2020. Using item co-occurrence statistics, we predicted the initial workup orders that would be entered by specialists and assessed the recommender's performance in a holdout data set based on what the specialists actually ordered. We surveyed endocrinologists to assess the clinical appropriateness of the predicted orders and to understand the initial workup process.

Results: Specialists (n=12) indicated that <50% of new patient referrals arrive with complete initial workup for common referral reasons. The algorithm achieved an area under the receiver operating characteristic curve of 0.95 (95% CI 0.95-0.96). Compared to a reference benchmark using the most common orders, precision and recall improved from 37% to 48% (*P*<.001) and from 27% to 39% (*P*<.001) for the top 4 recommendations, respectively. The top 4 recommendations generated for common referral conditions (abnormal thyroid studies, obesity, amenorrhea) were considered clinically appropriate the majority of the time by specialists surveyed and practice guidelines reviewed.

Conclusions: An item association–based recommender algorithm can predict appropriate specialists' workup orders with high discriminatory accuracy. This could support future clinical decision support tools to increase effectiveness and access to specialty referrals. Our study demonstrates important first steps toward a data-driven paradigm for outpatient specialty consultation with a tier of automated recommendations that proactively enable initial workup that would otherwise be delayed by awaiting an in-person visit.

XSL•FO
RenderX

Introduction

Background

There is a fundamental and growing gap between the supply and demand of medical expertise, as reflected in the projected shortage of 100,000 physicians by 2030 [1]. The problem is particularly acute for specialty care [2-6], for which over 25 million people in the United States have deficient access [7]. Wait times for in-person specialty visits commonly extend weeks to months after referrals are made [5]. Adding to this problem, essential initial workup is often not completed [8,9], resulting in ineffective visits when the specialists do not have sufficient information to make a definitive diagnosis and treatment recommendations by the time of their first in-person visit. Such inefficiency could lead to care delay, missed opportunity to provide access to more patients, and dissatisfaction of patients and families.

Ideally, referring providers could directly communicate with specialists for their preconsultation advice on an initial recommended clinical workup. However, data show that primary care providers are only able to communicate with specialists half of the time when referring patients [10]. Alternatively, primary care providers and specialists can collaboratively develop guidelines for initial workup [11], but this requires substantial manual effort to produce and maintain up-to-date content as new evidence arises and practice changes over time. Asynchronous electronic consults or synchronous telemedicine consults are emerging approaches for referring providers to solicit specialists' opinions on the need of referral and initial workup [12-16], with potential advantages of streamlining the referral process and empowering primary care providers. However, such consults remain limited in availability, as they still require a human consultant to review and respond to each request [17,18].

A more data-driven approach could boost the capacity of the health system by making initial specialty clinic visits more effective and by sparing the time required by specialists to communicate initial workup needs. Prior studies have shown the efficacy of statistical approaches, including association rules, Bayesian networks, logistic regression, and deep neural networks, for generating clinical order recommendations. The focus of these studies, however, has been primarily in the acute care settings such as inpatient hospitalization and emergency room visits [19-26].

Our aim is to develop a data-driven paradigm for outpatient specialty consultation with a tier of automated recommendations that proactively enable initial workup that would otherwise be delayed by awaiting an in-person visit. Taking advantage of electronic health records that contain thousands of specialist referral visits, we propose a data-driven algorithm inspired by Amazon's "customers who bought A also bought B" [27] to anticipate initial specialty evaluations at the time of referral based on how specialists cared for similar patients in the past. In this study, we chose pediatric endocrinology as a use case because laboratory evaluation is often required to inform specialist treatment recommendations [28-30].

Objective

Using specialty referrals to pediatric endocrinology as an example, we developed a recommender algorithm to anticipate initial workup needs for a variety of endocrine conditions. We compared the performance of the algorithm to a reference benchmark based upon the most common workup orders. We evaluated the need to complete initial workup and the clinical appropriateness of the algorithm recommendations by surveying specialists.

Methods

Recommender Algorithm Development

Deidentified structured electronic health record data from outpatient clinic visits at Stanford Children's Health were extracted from the Stanford Medicine Research Data Repository using the Observational Medical Outcomes Partnership (OMOP) common data model [31]. We include patients younger than 18 years with a pediatric endocrine referral order from any Stanford-affiliated clinic and a subsequent pediatric endocrine visit within 6 months. Between 2015 and 2020, 3424 patients met criteria, whose data yielded >1,150,000 instances of 8263 distinct clinical items.

We used OMOP common data model concepts to define distinct clinical items, including 2966 conditions, 2423 measurements (eg, lab results), 1187 procedures (eg, diagnostic imaging), and 1687 medications. Numeric laboratory results were categorized as "normal," "high," or "low" based upon reference ranges. We excluded clinical items that occurred in fewer than 10 patients.

Based on the timing of pediatric endocrinology referral, we split the patient cohort into a training set (referrals from 2015 to 2019: n=2842 patients) and a test set (referral in 2020: n=582 patients). In the training set, we calculated the co-occurrence statistics of pairs of clinical items to build an item association matrix (Figure 1). We counted duplicate items only once per patient to allow natural interpretation of patient prevalence and diagnostic measures.

XSL•FO
RenderX

Figure 1. Algorithm training and construction of the item co-occurrence matrix.
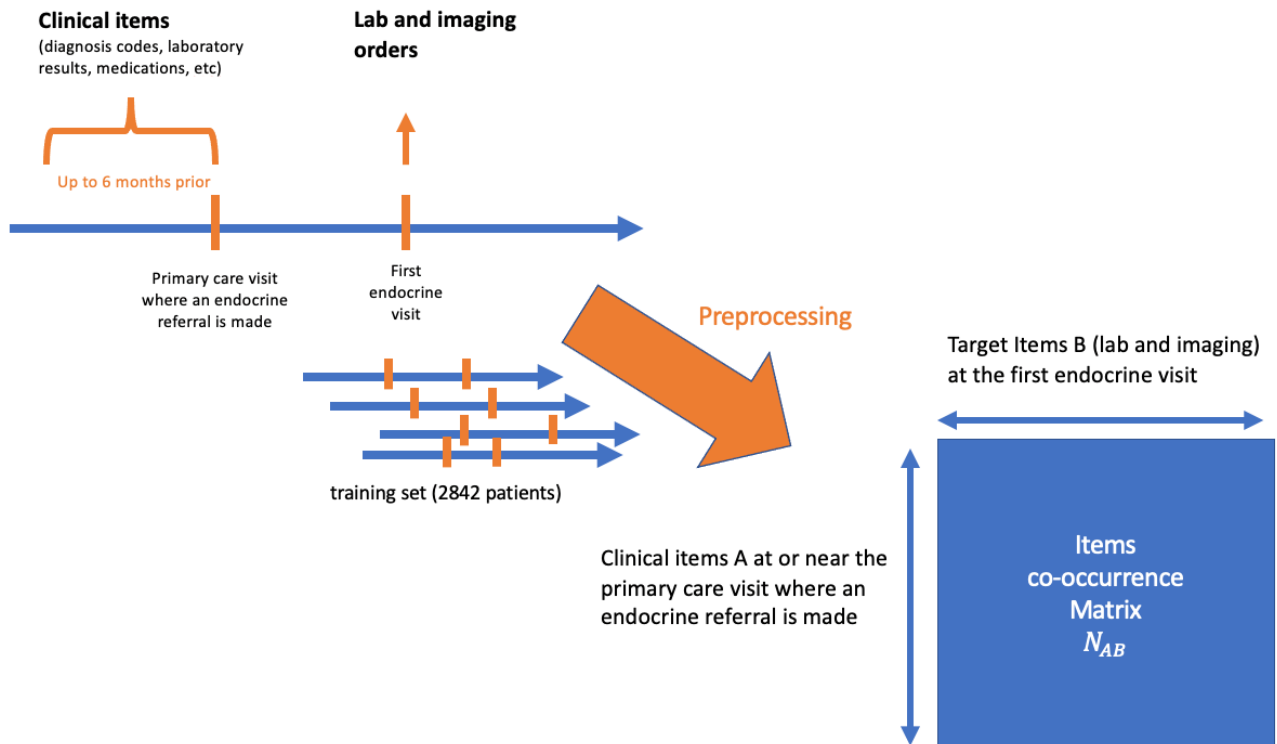
The recommender algorithm is queried with a patient's clinical items (diagnosis, labs, medications, etc) associated with the primary care encounter when the endocrine referral was placed. In addition, we included clinical items associated with the patient in the 6 months prior to the primary care encounter.

Using these clinical items (A_1,..., A_q), the recommender algorithm retrieves scores that resemble posttest probability from the co-occurrence association matrix for all possible target items at the subsequent endocrine visits. We limited the target items to laboratory and imaging orders to focus on diagnostic workup recommendations. For each query item (A), target items (B) are ranked by estimated posttest probability P(B|A), or positive predictive value (PPV), defined as the number of patients who have query item A followed by target item B (N_{AB}) divided by the number of patients with query item A (N_A).

If a patient has q query items, q separate ranked lists are generated. To aggregate these results, we estimate total pseudo-counts using the following equation that sums across every i-th query item:

W_A is a weighting factor for the query item. There are several ways one can model W_A. For instance, one can penalize common query items by the following expression:

Another method, inspired by a weighting strategy using item clustering based on genres [32], is to weigh a query item based

on its relevance to the endocrine referral cohort by using a relative risk term:

$$W_A = RR_A$$

Where:

The numerator is the prevalence of item A in our endocrine cohort ($N_{endocrine}$ is the total number of patients in our endocrine referral cohort, of which N_A patients have clinical item A). The denominator is the prevalence of item A outside of the endocrine cohort in all outpatient clinics (N_{outpt} is the total number of pediatric patients in all outpatient clinics, of which patients have item A).

Using 10-fold cross-validation in our training set, we evaluated these two strategies to model W_A individually and in combination (). We selected the W_A that gave the best prediction performance in the training data and subsequently used it in the test set.

The code can be accessed via GitHub [33].

Evaluation Using Electronic Health Record Test Set Data

To evaluate the performance of the recommender algorithm in the test set (Figure 2), we compared the recommended list of orders with the actual workup orders patients received at their first endocrine visit. We calculated the precision (PPV) and recall (sensitivity) for the top 4 recommendations, and performed the receiver operating characteristics analysis. We chose the top 4 recommendations because 4 is the mean number of workup

orders at the first endocrine visit. We calculated 95% CIs using   1000 bootstrap resamples [34].

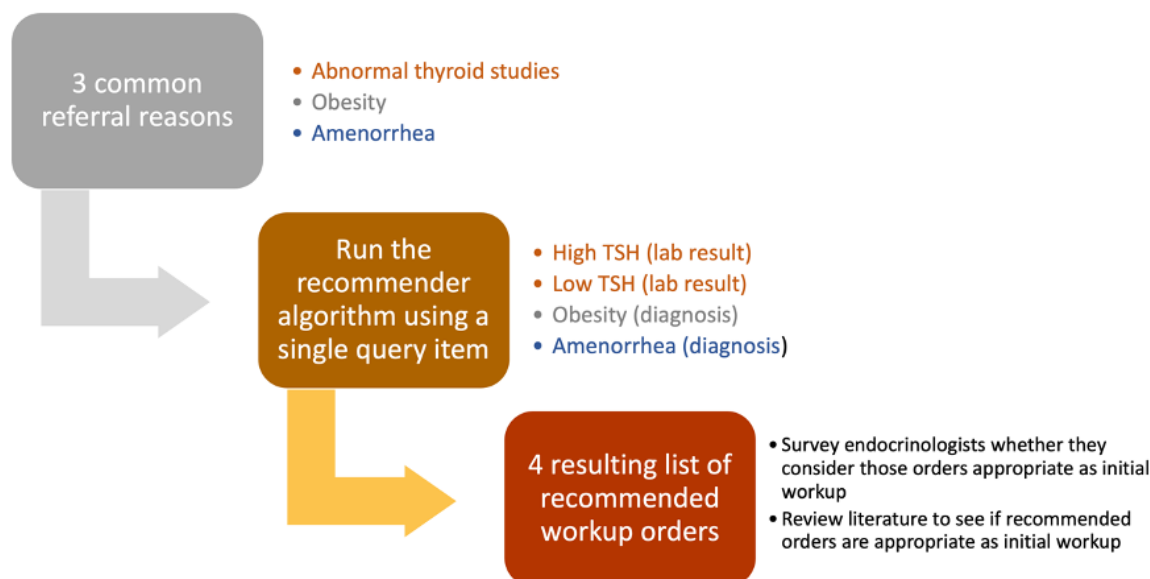**Figure 2.** Algorithm evaluation using the test set.
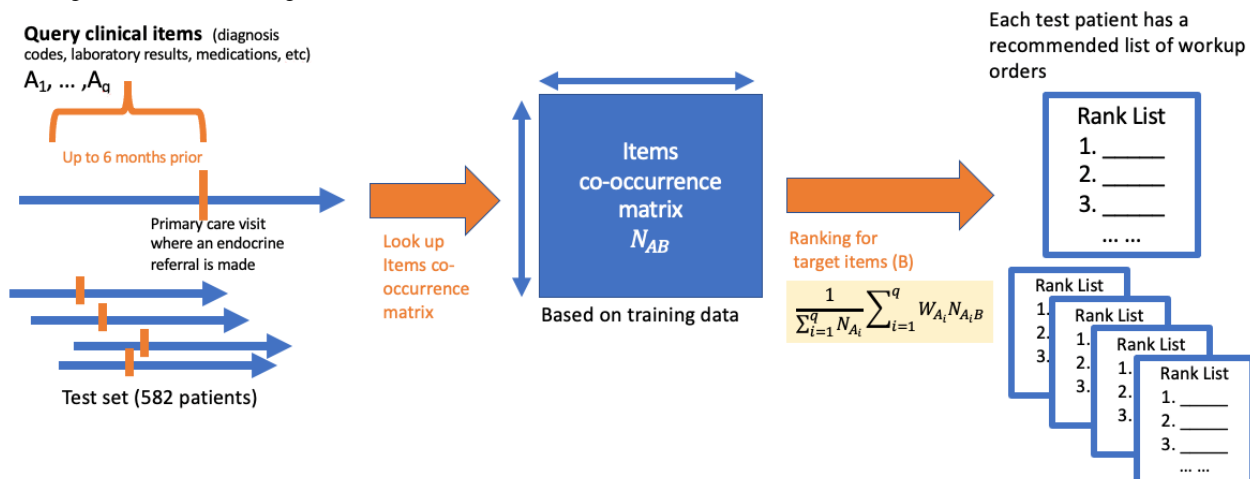
Evaluation Using Expert Surveys

To further understand whether the recommendations would be as clinically appropriate as the initial workup ordered by referring providers, we conducted a survey of all pediatric endocrinologists at Stanford Children's Health on three common referral reasons (abnormal thyroid studies, obesity, and amenorrhea). The survey was approved by the Institutional Review Board at Stanford University. Survey invitations were sent via emails in July 2020, and survey questions and informed consent were included in the supplemental material. For abnormal thyroid studies, we generated two lists of the top recommended workup orders—one queried with high thyroid stimulating hormone (TSH; an abnormal lab result suggesting hypothyroidism) and the other queried with low TSH (an abnormal lab result suggesting hyperthyroidism). For obesity

and amenorrhea, we generated a list of the top recommended orders using the diagnosis as a single query item (Figure 3). Subsequently, we asked the endocrinologists to select the orders from the recommended lists that they considered clinically appropriate as initial workup for the corresponding condition. Other than the referral reasons, the endocrinologists received no other information related to the patients. We instructed them to define appropriate workup as workup that gives sufficient information for the endocrinologists to make concrete recommendations at the clinic visits. In addition, for each of the three conditions, we asked them how often initial workup is completed in their practice and how helpful it is if initial workup is completed prior to the first specialty visit. Lastly, we reviewed published literature and consensus guidelines [28,30,35-37] as external validation to assess whether the recommended orders represent a reasonable workup.

Figure 3. Evaluation of algorithm output by practicing endocrinologists. TSH: thyroid stimulating hormone.

XSL•FO
RenderX

Results

Evaluation Using Electronic Health Record Test Set Data

Table 1 compares the performance of the recommender algorithm with a reference benchmark using the most common orders in our endocrine referral cohort (endocrine prevalence).

The recommender algorithm had the best performance with an area under the receiver operating characteristic curve (AUC) of 0.95 (95% CI 0.95-0.96). Comparing with the reference benchmark, precision improved from 37% to 48% (*P*<.001), and recall improved from 27% to 39% (*P*<.001). The recommender algorithm is based on a weighting factor, , that resulted in the best cross-validation performance in our training data (Multimedia Appendix 1, Table S1).

Table 1. Recommender algorithm performance in the test set. Precision and recall were calculated at k=4, given that 4 is the average number of workup orders.

	Recommender[a]	Endocrine prevalence[b]	Outpatient prevalence[c]	Random[d]
Precision[e] (%; 95% CI)	48 (45-52)	37 (34-40)	10 (8-12)	2 (2-3)
Recall[f] (%; 95% CI)	39 (36-42)	27 (24-29)	5 (4-6)	2 (1-3)
AUC[g] (95% CI)	0.95 (0.95-0.96)	0.88 (0.87-0.89)	0.64 (0.62-0.66)	0.49 (0.47-0.5)

[a]Recommender: ranking workup orders using the recommender algorithm.

[b]Endocrine prevalence: ranking workup orders using the percentage of patients who had the orders in the endocrine referral cohort () training set.

[c]Outpatient prevalence: ranking workup orders using the percentage of patients who had the orders among all outpatients ().

[d]Random: random ranking of workup orders.

[e]Precision: positive predictive value (proportion of predictions that were correct).

[f]Recall: sensitivity (proportion of correct items that were predicted).

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

Evaluation Using Expert Surveys

Of 14 pediatric endocrinologists at Stanford Children's Health, 12 (86%) responded to our survey on three common referral reasons (abnormal thyroid studies, obesity, and amenorrhea). Table 2 shows less than half of the patients coming to the specialty clinics with appropriate initial workup as estimated by the pediatric endocrinologists for each of the three referral reasons (each endocrinologist provided a value between 0% and 100%). The endocrinologists considered it as moderately to very helpful to have the initial workup completed prior to specialty visits (Table 2).

Table 3 shows the top recommendations based on the recommender algorithm using a single query item as mentioned in Figure 3. Each recommended workup order has a corresponding survey result showing the percentage of endocrinologists who considered the order clinically appropriate as the initial workup. Overall, the majority of the specialists considered the top four recommendations clinically appropriate in each of the lists.

Table 2. Estimated percentages of patients with initial workup completed before specialty visits and mean Likert scale score of how helpful (5: extremely helpful; 1: not helpful at all) it is to have initial workup completed before specialty visits.

	Value, mean (SD)
Estimated percentages of patients with initial workup completed before specialty visits (%)	
Abnormal thyroid studies	49 (21)
Obesity	45 (20)
Amenorrhea	37 (18)
Likert scale score of how helpful it is to have initial workup completed before specialty visits	
Abnormal thyroid studies	4.2 (0.8)
Obesity	3.3 (0.9)
Amenorrhea	4.1 (0.8)

Table 3. Top recommendations for patients referred for high thyroid stimulating hormone (TSH; commonly due to hypothyroidism), low TSH (commonly due to hypothyroidism), obesity, and amenorrhea.

Orders	PPV[a] (%)	Relative ratio[b]	Endocrine prevalence[c] (%)	Outpatient prevalence[d] (%)	Percent of endocrinologist considered appropriate
High TSH[e]					
TSH[f]	60.9	1.0	61.7	17.1	92
Free thyroxine[f]	60.3	1.2	56.8	7.5	100
Thyroglobulin antibody[f]	41.7	3.7	12.8	0.5	92
Thyroperoxidase antibody[f]	39.1	3.2	13.7	1.0	92
Vitamin D level	9.3	0.3	31.4	8.1	0
Serum cortisol	8.6	0.7	11.3	0.7	0
Low TSH[g]					
TSH[h]	61.5	1.0	61.7	17.1	75
Free thyroxine[h]	57.7	1.0	56.8	7.5	100
Thyroglobulin antibody[h]	50.0	4.0	12.8	0.5	67
Thyroperoxidase antibody[h]	46.2	3.4	13.7	1.0	58
Total tri-iodothyronine[h]	42.3	16.1	3.0	0.7	92
Comprehensive metabolic panel	26.9	0.5	53.8	23.1	8
Obesity[i]					
Hemoglobin A_{1c}[j]	40.2	1.9	22.5	12.5	100
TSH	28.0	0.4	61.7	17.1	75
Free thyroxine	25.6	0.4	56.8	7.5	42
Comprehensive metabolic panel[j]	25.6	0.5	53.8	23.1	92
Lipid panel[j]	25.0	1.4	18.3	12.9	100
Vitamin D level	20.7	0.6	31.4	8.1	42
Amenorrhea[k]					
Prolactin[l]	41.4	4.8	8.9	0.4	92
Luteinizing hormone[l]	37.9	2.2	17.5	0.9	100
Follicle stimulating hormone[l]	34.5	2.1	16.5	1.7	100
Estradiol	27.6	4.7	6.1	0.4	100
17 Hydroxy-progesterone	24.1	2.5	9.7	0.2	100
Dehydroepi-androsterone sulfate	17.2	2.2	8.0	0.4	92

[a]PPV: positive predictive value.

[b]Relative ratio: the ratio of the probability of the order given the query item to the probability of the order given the lack of the query item.

[c]Endocrine prevalence: the percentage of patients who had the orders in the endocrine referral cohort ( ).

[d]Outpatient prevalence: the percentage of patients who had the orders among all outpatients ( ).

[e]The top four orders are considered clinically appropriate by almost all of the endocrinologists and are recommended based on published guidelines. The fifth and sixth recommended items have relatively low PPV.

[f]Recommended based on guidelines [36].

[g]Here, the top five orders are considered clinically appropriate by most endocrinologists and published guidelines.

[h]Recommended based on guidelines [37].

[i]Hemoglobin A_{1c}, lipid panel, and comprehensive metabolic panel are considered clinically appropriate both by the endocrinologists and published guidelines.

[j]Recommended based on guidelines [35].

[k]The top six orders are considered clinically appropriate by almost all of the endocrinologists. The top three are also recommended based on published literature.

[l]Recommended based on published literature [30].

Discussion

Significance

Using pediatric endocrinology as an example, we developed and evaluated a recommender algorithm to anticipate initial workup needs at the time of specialty referral. The algorithm can predict appropriate specialist's workup orders with high discriminatory accuracy with an AUC>0.9. Our survey shows that, among the three common referral reasons, less than half of the patients typically have appropriate initial workup prior to their initial specialist visit. Most specialists agree that having initial workup completed prior to the first clinic visit is helpful and that our algorithm recommendations for the three referral conditions are clinically appropriate. This supports the potential utility of a data-driven recommender algorithm for referring providers. Although we illustrated 3 common referral conditions in this study, the algorithmic approach is general, and it could be broadly applied to other referral reasons or other specialties, with the benefit of personalization based on individual patient patterns of clinical items, including the combination of multiple conditions.

Although this algorithm is not suitable for full automation given the level of precision and recall, such an algorithm could serve as a clinical decision support tool [38-41] by displaying relevant clinical orders for referring providers to make the referral process more effective. One can imagine coupling this clinical decision support tool with electronic consultation so that specialists can quickly confirm the workup orders in the recommended list, thus augmenting the efficiency of the specialists and increasing their capacity to care for more patients. Advantages of an algorithmic decision support tool compared to building consensus guidelines among specialists [11,42] include scalability to answer unlimited queries on demand, maintainability through automated statistical learning, adaptability to respond to evolving clinical practices [43], and personalizability of individual suggestions with greater accuracy than manually authored checklists [43-45].

Different from our prior recommender algorithm for the inpatient setting [19], we applied a weighting factor to each query item based on its relevance to a specialty and its inverse frequency. The motivation is that inpatient clinical items are often related to acute reasons of hospitalization, while outpatient clinical items vary in scope, ranging from health maintenance or chronic disease management to treatment of urgent care issues. We show that differentially weighting query items significantly improves the performance of the recommender algorithm in both precision and recall. This makes intuitive sense because common clinical items seen in primary care clinics that are irrelevant to endocrinology likely provide less predictive power. A similar weighting scheme could be applied

to other recommender algorithms when the clinical use case is specialty specific.

The association rule mining methods shown here are relatively simple to implement with interpretable results and associated statistics. Other approaches including Bayesian networks [21] and deep machine learning [46] are computationally more complex with less interpretable results. Although future research should compare these different methods, our focus primarily is to demonstrate the applicability of a data-driven approach in workup recommendations for specialty referral.

Although we ranked the recommended workup items based on PPV as shown in Table 3, we have also provided alternative metrics such as relative ratio, which could be used to look for less common but more specific or "interesting" items for a given query. For instance, in Table 3, total tri-iodothyronine had a relative ratio of 16.1, suggesting this is highly specific for patients with low TSH (indicating hyperthyroidism). In comparison, free thyroxine ranks higher based on its PPV but has a relative ratio of 1.0, suggesting this is not specific for patients with low TSH. Indeed, we observed free thyroxine also appeared in the list of recommendations for patients with high TSH (Table 3).

For a crowdsourcing clinical decision support solution like recommender algorithms, a typical concern is that recommendations drawn from common practices do not necessarily imply clinical appropriateness. To address this concern, we solicited specialist opinions on the algorithm outputs. Overall, the majority of the top recommendations were considered clinically appropriate as initial workup by the specialists. We also performed external validation by reviewing relevant guidelines, which revealed general agreement with the specialists' assessments.

Limitations

Limitations in this study include that the algorithm was developed at a single institution, requiring future work to expand to other institutions to evaluate generalizability. However, the algorithmic framework is general, as we used a common data model with data schema and features that were not institution specific. Second, in recommender systems such as ours, there is a well-known cold start problem when there is a lack of clinical items. Our algorithm starts with a generic "best seller list" by using the cohort item prevalence, but the algorithm could rapidly bootstrap itself with even just a couple of clinical items such as diagnosis codes or laboratory results to dynamically converge on recommendations specific to the patient scenario. Third, our cohort definition relied on referral orders placed in the electronic health records, potentially failing to capture patients who were referred to specialty clinics by other means (eg, fax or phone communication). Additionally,

structured data in the electronic health records such as diagnosis codes or problem lists are often optimized for billing purpose and may be incomplete. Future research should investigate whether using unstructured text and leveraging natural language processing in clinical notes could further optimize the algorithm performance [47]. Fourth, our survey results are limited to three common referral conditions; further validation on other less common clinical conditions with more specialists from other institutions are needed. Future work should also include a prospective study to assess the effectiveness of the recommender algorithm in the specialty referral workflow. Lastly, this study did not include an analysis on the potential cost benefits of this recommender algorithm. Future research should compare the cost of additional visits due to incomplete workup with the cost of unnecessary labs if ordered based on algorithm recommendations.

Conclusion

An item association–based recommender algorithm can predict appropriate specialist's workup orders with high discriminatory accuracy. This could support future clinical decision support tools to increase effectiveness and access to specialty referrals. Our study demonstrates important first steps toward a data-driven paradigm for outpatient specialty consultation with a tier of automated recommendations that proactively enable initial workup that would otherwise be delayed by awaiting an in-person visit.

Authors' Contributions

WI conducted extraction and analysis of the data. WI and JHC have verified the underlying data. WI and PP conducted the survey study with specialists. WI drafted the manuscript. WI, PP, JP, and JHC contributed to the study concept and design, interpretation of data, and critical revision of the manuscript.

Conflicts of Interest

JHC is the cofounder of Reaction Explorer LLC that develops and licenses organic chemistry education software. He also received paid consulting or speaker fees from the National Institute of Drug Abuse Clinical Trials Network, Tuolc Inc, Roche Inc, and Younker Hyde MacFarlane PLLC. WI serves as Medical Director of Healthcare Data at nference and receives compensation. Other authors have no conflicts of interest to report.

Multimedia Appendix 1
Supplemental material and table.
[PDF File (Adobe PDF File), 215 KB - medinform_v10i3e30104_app1.pdf]

References

1. 2017 update: the complexities of physician supply and demand: projections from 2015 to 2030: final report. IHS Markit. URL: https://aamc-black.global.ssl.fastly.net/production/media/filer_public/a5/c3/a5c3d565-14ec-48fb-974b-99fafaeecb00/aamc_projections_update_2017.pdf [accessed 2020-12-04]
2. Mehrotra A, Forrest CB, Lin CY. Dropping the baton: specialty referrals in the United States. Milbank Q 2011 Mar;89(1):39-68 [FREE Full text] [doi: 10.1111/j.1468-0009.2011.00619.x] [Medline: 21418312]
3. Ray KN, Bogen DL, Bertolet M, Forrest CB, Mehrotra A. Supply and utilization of pediatric subspecialists in the United States. Pediatrics 2014 Jun;133(6):1061-1069. [doi: 10.1542/peds.2013-3466] [Medline: 24799548]
4. Jaakkimainen L, Glazier R, Barnsley J, Salkeld E, Lu H, Tu K. Waiting to see the specialist: patient and provider characteristics of wait times from primary to specialty care. BMC Fam Pract 2014 Jan 25;15:16 [FREE Full text] [doi: 10.1186/1471-2296-15-16] [Medline: 24460619]
5. Bisgaier J, Rhodes KV. Auditing access to specialty care for children with public insurance. N Engl J Med 2011 Jun 16;364(24):2324-2333. [doi: 10.1056/NEJMsa1013285] [Medline: 21675891]
6. Mayer ML. Are we there yet? Distance to care and relative supply among pediatric medical subspecialties. Pediatrics 2006 Dec;118(6):2313-2321. [doi: 10.1542/peds.2006-1570] [Medline: 17142513]

7.   Woolhandler S, Himmelstein DU. The relationship of health insurance and mortality: is lack of insurance deadly? Ann Intern Med 2017 Sep 19;167(6):424-431 [FREE Full text] [doi: 10.7326/M17-1403] [Medline: 28655034]

8.   Hendrickson CD, Saini S, Pothuloori A, Mecchella JN. Assessing referrals and improving information availability for consultations in an academic endocrinology clinic. Endocr Pract 2017 Feb;23(2):190-198. [doi: 10.4158/EP161514.OR] [Medline: 27849384]

9.   Hendrickson CD, Lacourciere SL, Zanetti CA, Donaldson PC, Larson RJ. Interventions to improve the quality of outpatient specialty referral requests: a systematic review. Am J Med Qual 2016 Sep;31(5):454-462. [doi: 10.1177/1062860615587741] [Medline: 26013165]

10. Stille CJ, McLaughlin TJ, Primack WA, Mazor KM, Wasserman RC. Determinants and impact of generalist-specialist communication about pediatric outpatient referrals. Pediatrics 2006 Oct;118(4):1341-1349. [doi: 10.1542/peds.2005-3010] [Medline: 17015522]

11. Ho CK, Boscardin CK, Gleason N, Collado D, Terdiman J, Terrault NA, et al. Optimizing the pre-referral workup for gastroenterology and hepatology specialty care: consensus using the Delphi method. J Eval Clin Pract 2016 Feb;22(1):46-52 [FREE Full text] [doi: 10.1111/jep.12429] [Medline: 26223584]

12. Chen AH, Murphy EJ, Yee HF. eReferral--a new model for integrated care. N Engl J Med 2013 Jun 27;368(26):2450-2453. [doi: 10.1056/NEJMp1215594] [Medline: 23802515]

13. Joschko J, Keely E, Grant R, Moroz I, Graveline M, Drimer N, et al. Electronic consultation services worldwide: environmental scan. J Med Internet Res 2018 Dec 21;20(12):e11112 [FREE Full text] [doi: 10.2196/11112] [Medline: 30578187]

14. Osman MA, Schick-Makaroff K, Thompson S, Bialy L, Featherstone R, Kurzawa J, et al. Barriers and facilitators for implementation of electronic consultations (eConsult) to enhance access to specialist care: a scoping review. BMJ Glob Health 2019;4(5):e001629 [FREE Full text] [doi: 10.1136/bmjgh-2019-001629] [Medline: 31565409]

15. Vimalananda VG, Orlander JD, Afable MK, Fincke BG, Solch AK, Rinne ST, et al. Electronic consultations (E-consults) and their outcomes: a systematic review. J Am Med Inform Assoc 2020 Mar 01;27(3):471-479 [FREE Full text] [doi: 10.1093/jamia/ocz185] [Medline: 31621847]

16. Kern C, Fu DJ, Kortuem K, Huemer J, Barker D, Davis A, et al. Implementation of a cloud-based referral platform in ophthalmology: making telemedicine services a reality in eye care. Br J Ophthalmol 2020 Mar;104(3):312-317 [FREE Full text] [doi: 10.1136/bjophthalmol-2019-314161] [Medline: 31320383]

17. Nabelsi V, Lévesque-Chouinard A, Liddy C, Dumas Pilon M. Improving the referral process, timeliness, effectiveness, and equity of access to specialist medical services through electronic consultation: pilot study. JMIR Med Inform 2019 Jul 10;7(3):e13354 [FREE Full text] [doi: 10.2196/13354] [Medline: 31293239]

18. Chang Y, Carsen S, Keely E, Liddy C, Kontio K, Smit K. Electronic consultation systems: impact on pediatric orthopaedic care. J Pediatr Orthop 2020 Oct;40(9):531-535. [doi: 10.1097/BPO.0000000000001607] [Medline: 32931692]

19. Chen JH, Podchiyska T, Altman R. OrderRex: clinical order decision support and outcome predictions by data-mining electronic medical records. J Am Med Inform Assoc 2016 Mar;23(2):339-348 [FREE Full text] [doi: 10.1093/jamia/ocv091] [Medline: 26198303]

20. Zhang Y, Padman R, Levin JE. Paving the COWpath: data-driven design of pediatric order sets. J Am Med Inform Assoc 2014 Oct;21(e2):e304-e311 [FREE Full text] [doi: 10.1136/amiajnl-2013-002316] [Medline: 24674844]

21. Klann JG, Szolovits P, Downs SM, Schadow G. Decision support from local data: creating adaptive order menus from past clinician behavior. J Biomed Inform 2014 Apr;48:84-93 [FREE Full text] [doi: 10.1016/j.jbi.2013.12.005] [Medline: 24355978]

22. Klann J, Schadow G, Downs S. A method to compute treatment suggestions from local order entry data. AMIA Annu Symp Proc 2010 Nov 13;2010:387-391 [FREE Full text] [Medline: 21347006]

23. Klann J, Schadow G, McCoy J. A recommendation algorithm for automating corollary order generation. AMIA Annu Symp Proc 2009 Nov 14;2009:333-337 [FREE Full text] [Medline: 20351875]

24. Hunter-Zinck HS, Peck J, Strout T, Gaehde SA. Predicting emergency department orders with multilabel machine learning techniques and simulating effects on length of stay. J Am Med Inform Assoc 2019 Dec 01;26(12):1427-1436 [FREE Full text] [doi: 10.1093/jamia/ocz171] [Medline: 31578568]

25. Wright A, Chen ES, Maloney FL. An automated technique for identifying associations between medications, laboratory results and problems. J Biomed Inform 2010 Dec;43(6):891-901 [FREE Full text] [doi: 10.1016/j.jbi.2010.09.009] [Medline: 20884377]

26. Wang JX, Sullivan D, Wells A, Chen JH. ClinicNet: machine learning for personalized clinical order set recommendations. JAMIA Open 2020 Jul;3(2):216-224 [FREE Full text] [doi: 10.1093/jamiaopen/ooaa021] [Medline: 32734162]

27. Smith B, Linden G. Two decades of recommender systems at Amazon.com. IEEE Internet Computing 2017 May;21(3):12-18. [doi: 10.1109/mic.2017.72]

28. Hanley P, Lord K, Bauer AJ. Thyroid disorders in children and adolescents: a review. JAMA Pediatr 2016 Oct 01;170(10):1008-1019. [doi: 10.1001/jamapediatrics.2016.0486] [Medline: 27571216]

29. Cuda SE, Censani M. Pediatric obesity algorithm: a practical approach to obesity diagnosis and management. Front Pediatr 2018;6:431. [doi: 10.3389/fped.2018.00431] [Medline: 30729102]

30. Klein DA, Poth MA. Amenorrhea: an approach to diagnosis and management. Am Fam Physician 2013 Jun 01;87(11):781-788 [FREE Full text] [Medline: 23939500]

31. Datta S, Posada J, Olson G, Li W, O'Reilly C, Balraj D, et al. A new paradigm for accelerating clinical data science at Stanford Medicine. arXiv. Preprint posted online on March 17, 2020 [FREE Full text]

32. Frémal S, Lecron F. Weighting strategies for a recommender system using item clustering based on genres. Expert Syst Applications 2017 Jul;77:105-113. [doi: 10.1016/j.eswa.2017.01.031]

33. HealthRex / CDSS. GitHub. URL: https://github.com/HealthRex/CDSS/tree/master/scripts/specialty_recommender [accessed 2022-02-22]

34. Davison AC, Hinkley DV. Bootstrap Methods and Their Application. Cambridge: Cambridge University Press; 1997.

35. Styne DM, Arslanian S, Connor E, Farooqi IS, Murad MH, Silverstein JH, et al. Pediatric obesity-assessment, treatment, and prevention: an Endocrine Society clinical practice guideline. J Clin Endocrinol Metab 2017 Mar 01;102(3):709-757 [FREE Full text] [doi: 10.1210/jc.2016-2573] [Medline: 28359099]

36. Child with suspected acquired hypothyroidism. Pediatric Endocrine Society. URL: https://pedsendo.org/clinical-resource/child-with-suspected-acquired-hypothyroidism/ [accessed 2021-04-16]

37. Child with suspected hyperthyroidism. Pediatric Endocrine Society. URL: https://pedsendo.org/clinical-resource/child-with-suspected-hyperthyroidism/ [accessed 2021-04-16]

38. Ostropolets A, Zhang L, Hripcsak G. A scoping review of clinical decision support tools that generate new knowledge to support decision making in real time. J Am Med Inform Assoc 2020 Dec 09;27(12):1968-1976 [FREE Full text] [doi: 10.1093/jamia/ocaa200] [Medline: 33120430]

39. Middleton B, Sittig DF, Wright A. Clinical decision support: a 25 year retrospective and a 25 year vision. Yearb Med Inform 2016 Aug 02;Suppl 1:S103-S116 [FREE Full text] [doi: 10.15265/IYS-2016-s034] [Medline: 27488402]

40. Sittig DF, Wright A, Osheroff JA, Middleton B, Teich JM, Ash JS, et al. Grand challenges in clinical decision support. J Biomed Inform 2008 Apr;41(2):387-392 [FREE Full text] [doi: 10.1016/j.jbi.2007.09.003] [Medline: 18029232]

41. Osheroff JA, Teich JM, Middleton B, Steen EB, Wright A, Detmer DE. A roadmap for national action on clinical decision support. J Am Med Inform Assoc 2007;14(2):141-145 [FREE Full text] [doi: 10.1197/jamia.M2334] [Medline: 17213487]

42. Cornell E, Chandhok L, Rubin K. Implementation of referral guidelines at the interface between pediatric primary and subspecialty care. Healthc (Amst) 2015 Jun;3(2):74-79. [doi: 10.1016/j.hjdsi.2015.02.003] [Medline: 26179727]

43. Chen JH, Alagappan M, Goldstein MK, Asch SM, Altman RB. Decaying relevance of clinical data towards future decisions in data-driven inpatient clinical order sets. Int J Med Inform 2017 Jun;102:71-79 [FREE Full text] [doi: 10.1016/j.ijmedinf.2017.03.006] [Medline: 28495350]

44. Chen JH, Goldstein M, Asch S, Mackey L, Altman RB. Predicting inpatient clinical order patterns with probabilistic topic models vs conventional order sets. J Am Med Inform Assoc 2017 May 01;24(3):472-480 [FREE Full text] [doi: 10.1093/jamia/ocw136] [Medline: 27655861]

45. Wang JK, Hom J, Balasubramanian S, Schuler A, Shah NH, Goldstein MK, et al. An evaluation of clinical order patterns machine-learned from clinician cohorts stratified by patient mortality outcomes. J Biomed Inform 2018 Oct;86:109-119 [FREE Full text] [doi: 10.1016/j.jbi.2018.09.005] [Medline: 30195660]

46. Islam MM, Yang H, Poly TN, Li YJ. Development of an artificial intelligence-based automated recommendation system for clinical laboratory tests: retrospective analysis of the National Health Insurance Database. JMIR Med Inform 2020 Nov 18;8(11):e24163 [FREE Full text] [doi: 10.2196/24163] [Medline: 33206057]

47. Lee H, Kang J, Yeo J. Medical specialty recommendations by an artificial intelligence chatbot on a smartphone: development and deployment. J Med Internet Res 2021 May 06;23(5):e27460 [FREE Full text] [doi: 10.2196/27460] [Medline: 33882012]

Abbreviations

AUC: area under the receiver operating characteristic curve
OMOP: Observational Medical Outcomes Partnership
PPV: positive predictive value
STARR: Stanford Medicine Research Data Repository
TSH: thyroid stimulating hormone

XSL•FO
RenderX

<u>Original Paper</u>

Predicting High Flow Nasal Cannula Failure in an Intensive Care Unit Using a Recurrent Neural Network With Transfer Learning and Input Data Perseveration: Retrospective Analysis

George Pappy[1*], MSc; Melissa Aczon[1*], PhD; Randall Wetzel[1], MBBS; David Ledbetter[1*], BSc

The Laura P. and Leland K. Whittier Virtual PICU, Children's Hospital Los Angeles, Los Angeles, CA, United States

[*] these authors contributed equally

Corresponding Author:
David Ledbetter, BSc
The Laura P. and Leland K. Whittier Virtual PICU
Children's Hospital Los Angeles
4650 Sunset Blvd
Los Angeles, CA, 90027
United States
Phone: 1 323 717 4515
Email: dledbetter@chla.usc.edu

Abstract

Background: High flow nasal cannula (HFNC) provides noninvasive respiratory support for children who are critically ill who may tolerate it more readily than other noninvasive ventilation (NIV) techniques such as bilevel positive airway pressure and continuous positive airway pressure. Moreover, HFNC may preclude the need for mechanical ventilation (intubation). Nevertheless, NIV or intubation may ultimately be necessary for certain patients. Timely prediction of HFNC failure can provide an indication for increasing respiratory support.

Objective: The aim of this study is to develop and compare machine learning (ML) models to predict HFNC failure.

Methods: A retrospective study was conducted using the Virtual Pediatric Intensive Care Unit database of electronic medical records of patients admitted to a tertiary pediatric intensive care unit between January 2010 and February 2020. Patients aged <19 years, without apnea, and receiving HFNC treatment were included. A long short-term memory (LSTM) model using 517 variables (vital signs, laboratory data, and other clinical parameters) was trained to generate a continuous prediction of HFNC failure, defined as escalation to NIV or intubation within 24 hours of HFNC initiation. For comparison, 7 other models were trained: a logistic regression (LR) using the same 517 variables, another LR using only 14 variables, and 5 additional LSTM-based models using the same 517 variables as the first LSTM model and incorporating additional ML techniques (transfer learning, input perseveration, and ensembling). Performance was assessed using the area under the receiver operating characteristic (AUROC) curve at various times following HFNC initiation. The sensitivity, specificity, and positive and negative predictive values of predictions at 2 hours after HFNC initiation were also evaluated. These metrics were also computed for a cohort with primarily respiratory diagnoses.

Results: A total of 834 HFNC trials (455 [54.6%] training, 173 [20.7%] validation, and 206 [24.7%] test) met the inclusion criteria, of which 175 (21%; training: 103/455, 22.6%; validation: 30/173, 17.3%; test: 42/206, 20.4%) escalated to NIV or intubation. The LSTM models trained with transfer learning generally performed better than the LR models, with the best LSTM model achieving an AUROC of 0.78 versus 0.66 for the 14-variable LR and 0.71 for the 517-variable LR 2 hours after initiation. All models except for the 14-variable LR achieved higher AUROCs in the respiratory cohort than in the general intensive care unit population.

Conclusions: ML models trained using electronic medical record data were able to identify children at risk of HFNC failure within 24 hours of initiation. LSTM models that incorporated transfer learning, input data perseveration, and ensembling showed improved performance compared with the LR and standard LSTM models.

Introduction

Background

The use of high flow nasal cannula (HFNC) respiratory support in children in critical care, emergency departments, and general wards has increased in recent years [1-8]. HFNC provides an alternative to other noninvasive ventilation (NIV) techniques and endotracheal intubation, has fewer associated risks and complications, and is well-tolerated by children [3,5,6,8]. Nevertheless, many patients require escalation to a higher level of respiratory support [3,8]. Importantly, for those who require escalation, recent research indicates better clinical outcomes for patients who were escalated to higher levels of respiratory support earlier: lower hospital and intensive care unit (ICU) mortality rates, higher extubation success rate, higher ventilator-free days, and lower hospital and ICU lengths of stay [9,10]. These findings suggest that early identification of children in whom HFNC will not be successful could allow more timely institutions of advanced respiratory support and decrease morbidity and mortality.

Goals

This study aims to develop a model to make reliable, real-time predictions of a child's response to HFNC. Such a model could help clinicians differentiate three groups: (1) children likely to do well on HFNC alone, (2) children likely to need a higher level of support, and (3) children whose HFNC response is unclear. Differentiating these 3 groups would help clinicians resolve the dilemma of appropriate NIV while not unduly and potentially harmfully prolonging it. The last group may benefit from the closest and most frequent monitoring. The second group, although still monitored frequently, could be escalated by clinicians to a higher level of support earlier. A further goal is to compare different algorithms, from logistic regressions (LRs) to long short-term memory (LSTM)-based recurrent neural networks, for predicting HFNC response. Other techniques, such as transfer learning (TL), input data perseveration, and ensembling, are also explored and evaluated for their impact on performance when used with LSTMs.

Related Prior Work

The authors are unaware of any studies developing a predictive model of HFNC failure, although a few studies have investigated risk factors for escalation from HFNC to a higher level of respiratory support. Guillot et al [11] found that high pCO_2 (partial pressure of carbon dioxide) was a risk factor for HFNC failure in children with bronchiolitis. Er et al [12] reported that respiratory acidosis, low initial oxygen saturation and SF (oxygen saturation [SpO_2] divided by the fraction of inspired oxygen [FiO2]) ratio, and SF ratio <195 during the first few hours were associated with unresponsiveness to HFNC in children with severe bacterial pneumonia in a pediatric emergency department. In a small study of children with bacterial pneumonia, Yurtseven and Saz [13] saw higher failure rates in those with higher respiratory rates. Kelley et al [8] found that a high respiratory rate, high initial venous pCO_2, and a pH <7.3 were associated with failure of HFNC.

Methods

Data Sources

Data for this study came from deidentified clinical observations collected in electronic medical records (EMRs; Cerner) of children admitted to the pediatric intensive care unit (PICU) of Children's Hospital Los Angeles (CHLA) between January 2010 and February 2020. An episode represents a single admission and a contiguous stay in the PICU. Patients may have >1 episode. EMR data for an episode included irregularly, sparsely, and asynchronously charted physiological measurements (eg, heart rate and blood pressure), laboratory results (eg, creatinine and glucose level), drugs (eg, epinephrine and furosemide), and interventions (eg, intubation, bilevel positive airway pressure [BiPAP], or HFNC). Data previously collected for Virtual Pediatric Services, LLC participation [14], including diagnoses, gender, race, and disposition at discharge, were linked with the EMR data before deidentification.

Ethics Exemption

The CHLA institutional review board reviewed the study protocol and waived the requirement for consent and institutional review board approval.

Definitions

For ease of reference, Textbox 1 lists the terminologies and definitions used throughout the sections which follow.

Textbox 1. Useful definitions.

Terms and definitions

- Episode: An individual child's single, contiguous stay in the pediatric intensive care unit, spanning the time between admission and discharge

- High flow nasal cannula (HFNC) initiation: The start of HFNC treatment for a child not currently on HFNC

- HFNC period: The 24 hours following an HFNC initiation where the child was not on HFNC support at any time during the preceding 24 hours

- HFNC trial: An episode or subset of an episode (starting with admission) where only the very last HFNC period is designated as the training target; it may include previous HFNC initiations

---

In an episode where HFNC is initiated only once, there is exactly 1 HFNC period and 1 HFNC trial. Episodes can have multiple HFNC initiations. In such cases, a single episode may have multiple HFNC periods, and each has an associated HFNC trial. Note that not all HFNC initiations have a corresponding HFNC period. For instance, if a child started on HFNC for the first time during an episode, then this marked the start of the HFNC period. If HFNC was withdrawn 2 hours later, and the child again received HFNC an hour after that, then this new HFNC initiation did not mark the start of a new HFNC period as the

original HFNC period had not yet ended. In contrast, if this second HFNC initiation took place >24 hours after the first HFNC was stopped, then this second initiation marked the start of a new HFNC period as it was initiated after the first HFNC period had already ended. Finally, at least 30 minutes was required between any de-escalation (*step-down*) from NIV or intubation before the start of the HFNC period. This rule was necessary as patients on a higher level of support may be stepped down to HFNC to assess their ability to breathe on their own. If such breathing trials fail, which is not an uncommon occurrence, these patients immediately escalate back to mechanical ventilation or NIV, technically becoming HFNC failures but were, in fact, extubation failures and are not representative of the escalation scenarios of interest in this study. Figure 1 illustrates these terminologies.

Figure 1. Illustration of HFNC scenarios, definitions, and outcomes. HFNC: high flow nasal cannula; ICU: intensive care unit; NIV: noninvasive ventilation.
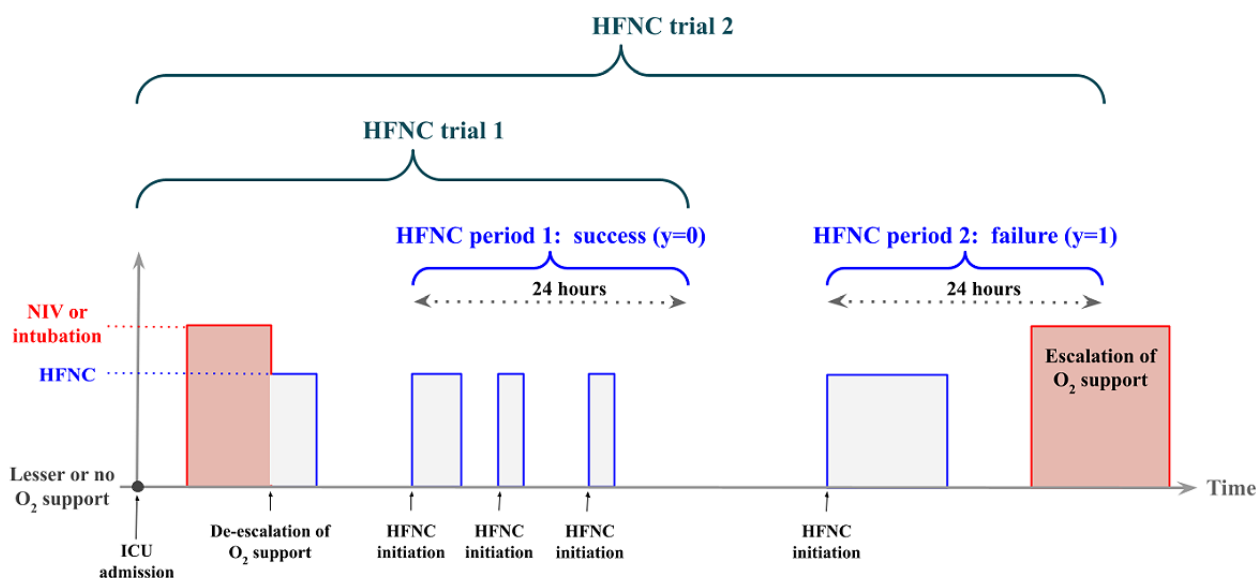
Data Inclusions and Exclusions

Only episodes in which HFNC was used were included. Episodes of patients aged ≥19 years at admission were excluded, as were episodes associated with sleep apnea. Any episode that ended <24 hours into an HFNC period where the patient next went to the operating room was also excluded. Episodes with a do not intubate or do not resuscitate order were also excluded.

Target Outcome

For each HFNC trial, the target of interest was escalation to a higher level of support (BiPAP, noninvasive mechanical ventilation, and intubation) within the 24-hour window (HFNC period) after HFNC initiation. Each HFNC trial was labeled either a failure (if there was an escalation within the associated HFNC period) or a success (if there was no such escalation within the associated HFNC period).

When a patient was discharged from the PICU within 24 hours of HFNC initiation, the target label was determined by the patient's disposition at discharge (Textbox 2). Episodes with the dispositions *operating room*, *another hospital's ICU*, or *another ICU in current hospital* were excluded as the outcome was ambiguous. HFNC trials associated with a favorable disposition (*general care floor*, *home*, or *step-down unit*) during the HFNC period were labeled as success.

Textbox 2. Target outcome mappings for high flow nasal cannula periods cut short by patient discharge.

Target outcome mapping and episode disposition

Success

- General care floor

- Home

- Step-down unit or intermediate care unit

Censored

- Operating room

- Another hospital's intensive care unit

- Another intensive care unit in current hospital

---

The HFNC definitions and outcomes, combined with the exclusion criteria, resulted in 834 HFNC trials that were randomly divided into training, validation, and hold-out test sets. All HFNC trials of an individual patient were assigned in only one of these 3 sets to prevent leakage and bias during model evaluations. No additional stratifications were applied.

Labeling Time Series Data for Model Training and Assessment

Recall that the task is to predict HFNC failure (escalation of care) for each HFNC trial. Data processing starts at the

beginning of the HFNC trial, with a model trained to output a prediction each time a new measurement becomes available. Figure 2 illustrates how the time series data of each HFNC trial were labeled for this process. All prediction times during the HFNC period were labeled as either 1 (failure) or 0 (success). Predictions at times before the HFNC period were labeled *NaN* (*Not a Number*) to exclude the predictions from error metrics during model training and performance evaluations.

Figure 2. Illustration of labeling time series data for predicting HFNC escalation. HFNC: high nasal flow cannula; NaN: Not a Number; PICU: pediatric intensive care unit.
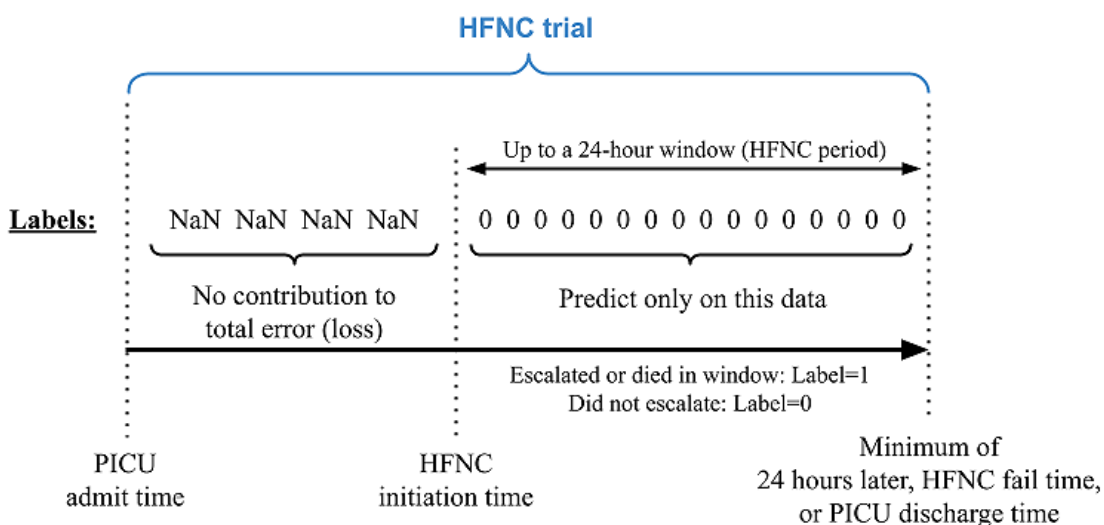
Data Preprocessing

Overview

Each episode's time series data were converted into a matrix. Rows contain the measurements (recorded or imputed) of all variables at 1 time point, and columns contain values of a single variable at different times. The steps of this conversion are described in detail in a previous work [15] and comprise the aggregation and normalization of observed measurements, followed by the imputation of missing data. A brief description is provided in the following sections.

Aggregation and Normalization

Where medically appropriate, values of the same variable obtained using independent measurement methods were aggregated into a single feature. For example, invasive and noninvasive systolic blood pressure measurements were grouped into a single variable representing the systolic blood pressure [16]. Any drug or intervention administered in <1% of patient episodes in the training set was excluded. This aggregation and exclusion process resulted in a list of 516 distinct demographic, physiological, laboratory, and therapy variables available as model inputs (see Tables S1 to S4 in Multimedia Appendices 1-4 for the full list; variable acronyms appear in Table S5 in Multimedia Appendix 5). Measurements considered

incompatible with human life were filtered out using established minimum and maximum acceptable values (eg, heart rates >400 beats per minute). Physiological variables and laboratory measurements were transformed to have 0 mean and unit variance using the means and SDs derived from the training set. Administered patient therapies were scaled to the interval [0,1] using clinically defined upper limits. No variables were normalized by age as patient age was one of the inputs. Diagnoses were only used for descriptive analyses and not as model input features.

Imputation

EMR measurements were sparsely, asynchronously, and irregularly charted, with time between measurements ranging from 1 minute to several hours. At any time where at least one variable had a recorded value, the missing values for other unrecorded variables were imputed. The imputation process depended on the type of variable. Missing measurements for a drug or an intervention variable were set to 0, indicating the absence of treatment. When physiological observations or laboratory measurements were available, they were propagated forward until another measurement was recorded. This choice reflects the clinical practice and is based on the observation that measurements are recorded more frequently when the patient is unstable and less frequently when the patient appears stable

[17]. If a physiological or laboratory variable had no recorded value throughout the entire episode, the mean from the training set population was used.

Input Perseveration

As described in the study by Ledbetter et al [18], LSTMs exhibit predictive lag, wherein the model fails to react quickly to new clinical information. A previous study [18] demonstrated that an LSTM trained with input data perseveration (ie, the input is replicated k times) responds with more pronounced changes in predictions when new measurements become available while maintaining overall performance relative to a standard LSTM. As timely model responsiveness to acute clinical events is critical in determining the necessity of escalating support, input data perseveration was assessed as a training augmentation technique.

Transfer Learning

TL is a technique of applying insights (eg, data representations) that were previously learned from one problem to a new, related task [19-22]. It can be particularly beneficial when one task has significantly more training data than the other. As the number of children on HFNC is significantly smaller than the number of ICU episodes in the CHLA PICU data set, TL techniques were considered to generate initial data representations and facilitate training of the HFNC prediction models. LSTM-based recurrent neural networks using the same input variables as those for the HFNC task were trained on >9000 CHLA PICU episodes to predict ICU mortality [23]. The first layer of one of these mortality models was then used as the first layer of some of the LSTM-based HFNC prediction models in Textbox 3.

Textbox 3. Details of the 8 models considered.

Model and hyperparameters (a list of the 14 variables used as model inputs appears in Table S6 of Multimedia Appendix 6)

14-variable logistic regression (LR-14)

- Regularizer: 7.50×10^{-1}

- Regularization: elasticnet (ratio=0.5)

517-variable logistic regression (LR-517)

- Regularizer: 1.15×10^{-3}

- Regularization: elasticnet (ratio=0.2)

Long short-term memory (LSTM)

- Layers: 3

- Number of hidden units: (128,256,128)

- Batch size: 12

- Initial learning rate: 9.6e-4

- Patience: 10

- Reduce rate: 0.9

- Number of rate reductions: 8

- Loss function: binary cross-entropy

- Optimizer: rmsprop

- Dropout: 0.35

- Recurrent dropout: 0.2

- Regularizer: 1e-4

- Output activation: sigmoid

LSTM with 3-times input perseveration (LSTM+3xPers)

- Same as LSTM

LSTM with transfer learning (TL; LSTM+TL)

- Same as LSTM

- Transfer weights: first hidden layer only

LSTM with 3-times input perseveration and TL (LSTM+3xPers+TL)

- Same as LSTM

- Transfer weights: first hidden layer only

Simple ensemble of LSTM+3xPers+TL (Simple-en-LSTM+3xPers+TL)

- Same as LSTM

- Transfer weights: first hidden layer only

Multi-ensemble of LSTM+3xPers+TL (Multi-EN-LSTM+3xPers+TL)

- Same as LSTM

- Transfer weights: first hidden layer only

Ensembling

Ensemble methods combine multiple algorithms to achieve a higher predictive performance than each component could obtain [24]. The predictions from each component are averaged to yield a single final prediction. Here, different seed values were used to generate multiple LSTM-based models, with each seed value initializing a different set of pseudorandom starting weights for a particular model. Different seed values led to slightly different models. Different seeds were used to train the

mortality models used for TL and train the final LSTM models on HFNC-specific data. Owing to the relatively small size of the cohort available to develop the HFNC prediction model, it was hypothesized that training models with different seeds would result in high variance and low bias models that may be decorrelated. Ensembling provides a method to average the results across decorrelated models to reduce variance but maintain a low bias.

HFNC Models

A total of 8 models were developed: a 14-variable LR (LR-14) using variables previously identified as risk factors for HFNC failure [8,11-13], a 517-variable LR (LR-517), a standard LSTM, an LSTM with input perseveration (LSTM+3xPers, where *3* indicates the number of replications described in the study by Ledbetter et al [18]), an LSTM with TL (LSTM+TL), an LSTM with both input perseveration and TL (LSTM+3xPers+TL), a simple ensemble of LSTMs with input perseveration and TL (Simple-EN-LSTM+3xPers+TL), and an ensemble of ensembles of LSTMs with input perseveration and TL (Multi-EN-LSTM+3xPers+TL). All models were trained to generate a prediction every time new measurements became available within the HFNC period.

Textbox 3 describes the parameters of all the models. Each model was developed on the training set to maximize the average of the validation set area under the receiver operating characteristic (AUROC) curves measured hourly from 0 to 14 hours into the HFNC period. This window was selected to prioritize the most clinically impactful period.

Figure 3 illustrates how the ensemble models were formed. An LSTM mortality model was trained (*seed A*), and its first layer was used as the first layer (TL weights) of a 3-layer LSTM with input perseveration. Layers 2 and 3 of this model were trained on the HFNC data 5 times (*seeds 1-5*), resulting in 5 slightly different HFNC models whose predictions were averaged to generate the Simple-EN-LSTM+3xPers+TL model predictions. This process was repeated 4 times to generate an ensemble of ensembles model: 4 LSTM mortality models were trained (*seeds A-D*), each providing a different set of TL weights. For each of these 4 sets of TL weights, 5 different seeds were used to train with the HFNC data, resulting in 20 models whose predictions were averaged together to generate the Multi-EN-LSTM+3xPers+TL model predictions.

Figure 3. Forming the (A) simple ensemble and (B) multi-ensemble models. HFNC: high flow nasal cannula; LSTM: long short-term memory; TL: transfer learning.
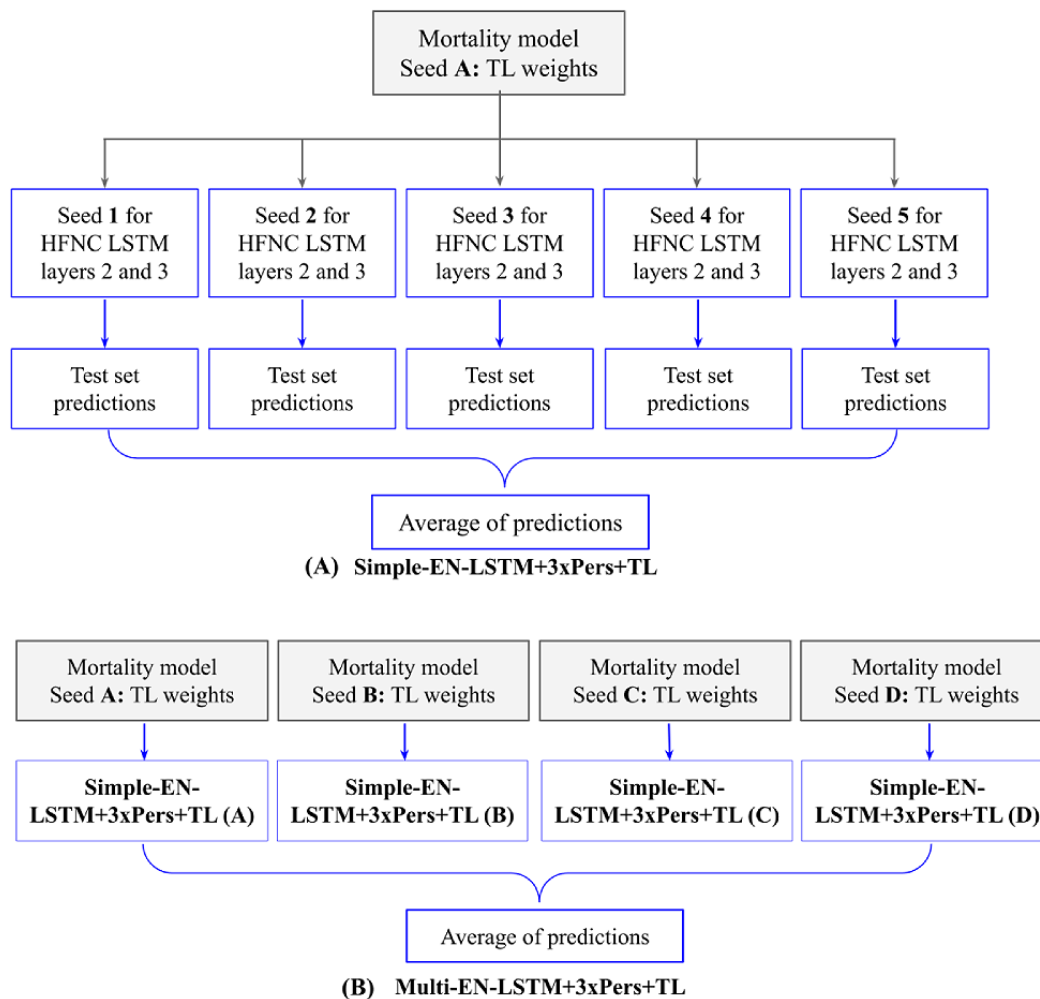
Model Performance Assessment

Model performance was assessed on the test set by evaluating the AUROC of predictions every 30 minutes within the 24-hour HFNC period. AUROC performance for the subset of patients with respiratory diagnoses was also compared every 30 minutes within the 24-hour HFNC period. In the *rolling cohort* AUROC computations, failures or successes that had already occurred before the time of evaluation were excluded. For example, any HFNC failures or successes that took place ≤4.5 hours into the HFNC period were not considered in computing the 5-hour AUROC. Including these in the 5-hour AUROC calculation would artificially boost the result. This was followed for all time points of interest. Therefore, the number of HFNC failures and successes in the test set steadily decreased from 0 to 24 hours in the HFNC period. *The fixed cohort* AUROC and area under the precision–recall curve in the first 15 hours were also computed, wherein only those who were on HFNC for at least 15 hours were included to ensure a constant cohort (and, consequently, a constant incidence rate of HFNC failures) at each evaluation point.

In addition, receiver operating characteristic (ROC) curves, sensitivities, specificities, positive predictive values (PPVs), and negative predictive values (NPVs) of predictions 2 hours after HFNC initiation were generated to evaluate model performance early in the HFNC period, on both the entire test set cohort and the respiratory subcohort.

Results

Cohort Characteristics

Table 1 describes the demographics and characteristics of the data, whereas Figure 4 shows the histogram (in cyan) and cumulative density (orange) for the time to HFNC failure for the entire data set. Approximately 50% (87/175) of failures occurred within 7.6 hours, and 80% (140/175) occurred within 14.1 hours.

Table 1. Demographics and characteristics of the data partitions (N=834).

Characteristic	Training set (n=455)	Validation set (n=173)	Test set (n=206)	Overall (n=834)
Patients, n	341	138	158	637
Episodes, n	381	151	183	715
HFNC[a] trials died, n (%)	21 (4.6)	10 (5.8)	7 (3.4)	38 (4.6)
HFNC trials failed, n (%)	103 (22.6)	30 (17.3)	42 (20.4)	175 (21)
HFNC trials female, n (%)	200 (44)	70 (40.5)	90 (43.7)	360 (43.2)
HFNC trials with respiratory primary diagnosis, n (%)	333 (73.2)	115 (66.5)	141 (68.4)	589 (70.6)
PRISM[b] 3 score				
Values, mean (SD)	4.4 (5.3)	3.6 (5.0)	4.0 (5.0)	4.2 (5.2)
Values, median (IQR)	3 (0-7)	2 (0-5)	2 (0-6)	3 (0-6)
Age (years)				
Values, mean (SD)	3.3 (4.6)	3.1 (4.5)	2.8 (3.7)	3.1 (4.4)
Values, median (IQR)	1.2 (0.4-3.4)	1.2 (0.5-3.1)	1.2 (0.5-3.8)	1.2 (0.4-3.5)
Age group (years), n (%)				
0-1	205 (45.1)	78 (45.1)	96 (46.6)	379 (45.4)
1-5	164 (36)	63 (36.4)	77 (37.4)	304 (36.5)
5-10	31 (6.8)	12 (6.9)	17 (8.3)	60 (7.2)
10-19	55 (12.1)	20 (11.6)	16 (7.8)	91 (10.9)

[a]HFNC: high flow nasal cannula.

[b]PRISM: pediatric risk of mortality.

Figure 4. Distribution of time to HFNC failure. HFNC: high flow nasal cannula.
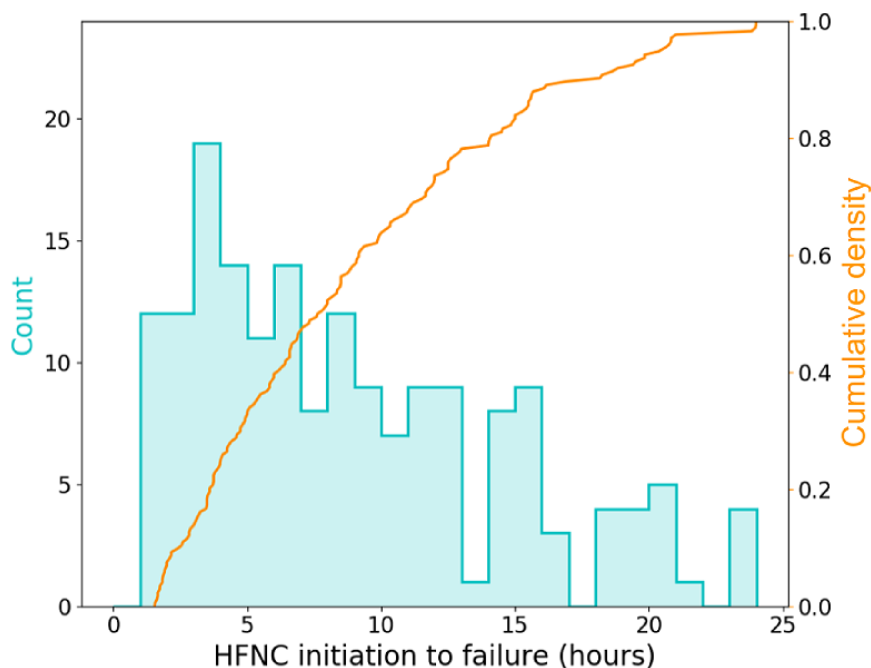

AUROC Across the First 24 Hours

Figure 5 shows the 8 models' *rolling cohort* AUROCs in 30-minute increments within the 24-hour HFNC period of all HFNC trials in the test set. Table S7 in Multimedia Appendix 7 shows the number of remaining HFNC trials in the test cohort at various evaluation times. Table S8 in Multimedia Appendix 8 presents the test set AUROC values associated with Figure 5

at several times of interest in the first 12 hours of the HFNC period. Table S9 in Multimedia Appendix 9 presents the corresponding AUROCs in the respiratory cohort. The *fixed cohort* AUROCs and areas under the precision–recall curve are shown in Figures S10 and S11 in Multimedia Appendices 10 and 11. Both the rolling and fixed cohort AUROCs generally increased over time.

Figure 5. Area under the receiver operating characteristics (AUROCs) of model predictions at different times on hold-out test set. AUC: area under the receiver operating characteristic curve; HFNC: high flow nasal cannula; LR: logistic regression; LSTM: long short-term memory; TL: transfer learning.
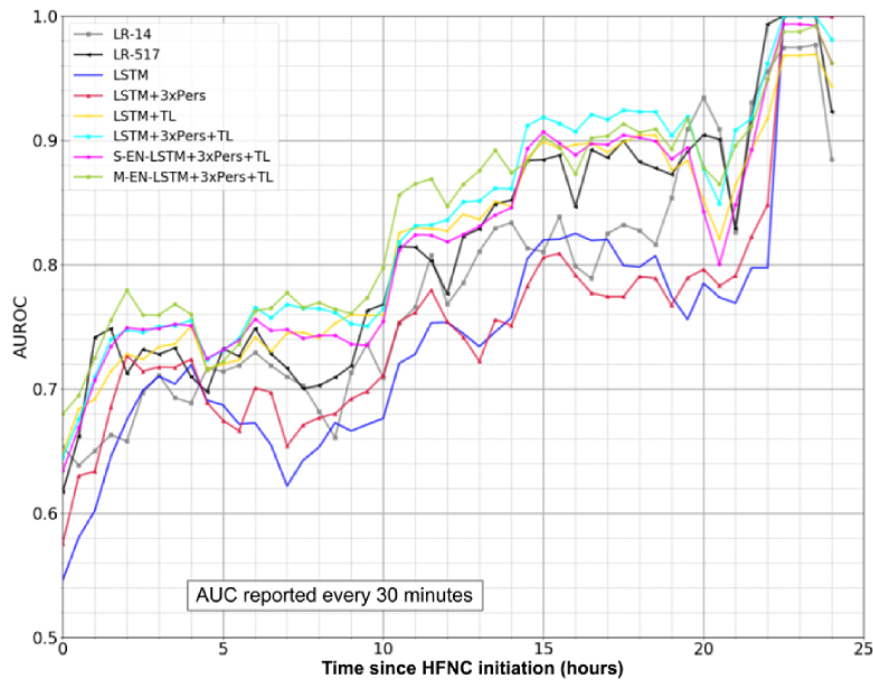

## Two-Hour ROC and AUROC

Figure 6 presents the test set 2-hour ROC curves and AUROC for the 8 models, showing predictive performance just 2 hours into the HFNC period, whereas Figure 7 presents the same metrics corresponding to the respiratory cohort.

Figure 6. Receiver operating characteristic curves and area under the receiver operating characteristic (AUROC) curves of 2-hour predictions on the entire test set. LR: logistic regression; LSTM: long short-term memory; TL: transfer learning.
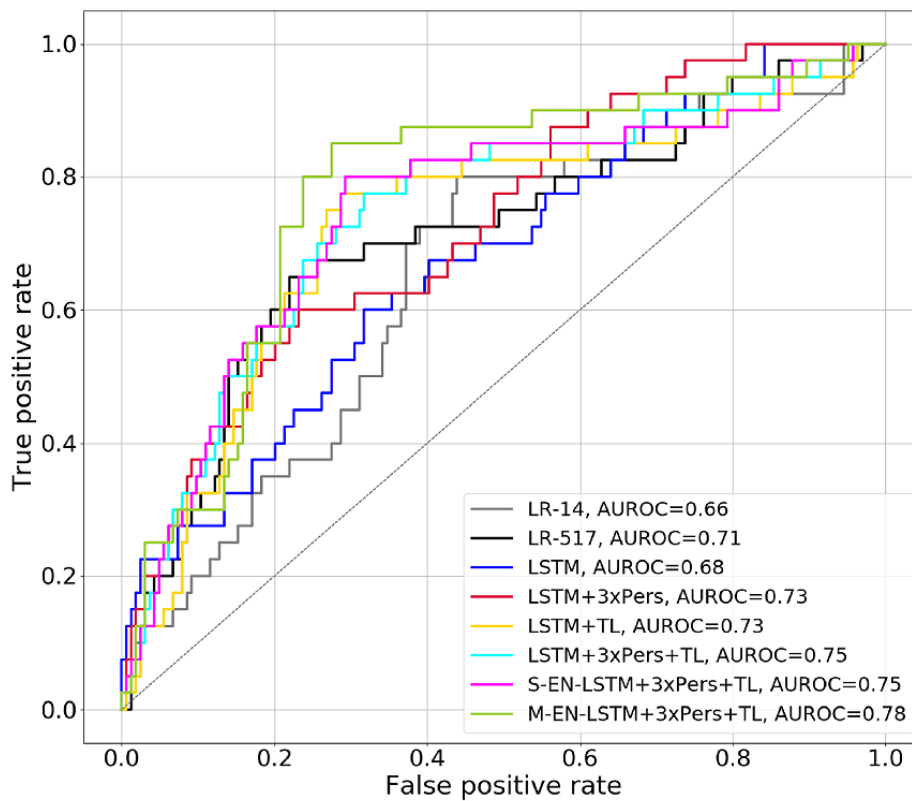
Figure 7. Receiver operating characteristic curves and area under the receiver operating characteristic (AUROC) curves of 2-hour predictions on high flow nasal cannula trials whose primary diagnosis is respiratory: all models. LR: logistic regression; LSTM: long short-term memory; TL: transfer learning.
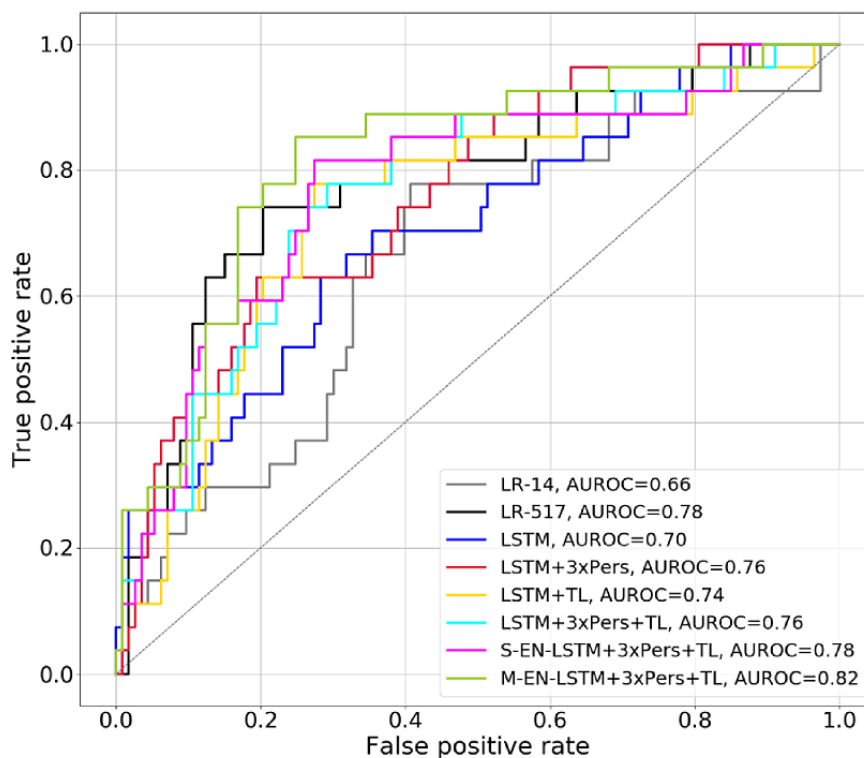
Positive Predictive Value and Negative Predictive Value

Table 2 shows the specificity, PPV, and NPV for the 2-hour predictions of the Multi-EN-LSTM+3xPers+TL model that correspond to different values of sensitivity. These metrics provide a more intuitive understanding of performance in a deployment scenario. At the 2-hour mark, 204 HFNC periods

remained (40 [19.6%] failures and 164 [80.4%] successes); setting the operating point at 25% sensitivity correctly identified 10 of the HFNC failures (PPV=67%) and 159 of the nonfailures (NPV=84%). Among the correctly identified HFNC failures, the time to failure (at the 2-hour mark) ranged from a few minutes to 19 hours (median=2.6 hours). Tables S12 to S14 in Multimedia Appendices 12-14 show a comparison of these metrics across all models.

Table 2. Specificity, PPV,[a] and NPV[b] corresponding to various sensitivity values of the 2-hour predictions of the Multi-EN-LSTM+3xPers+TL model.

Sensitivity	Specificity	PPV	NPV
0.10	0.982	0.571	0.817
0.20	0.970	0.615	0.832
0.25	0.970	0.667	0.841
0.30	0.927	0.500	0.844
0.40	0.848	0.390	0.853
0.50	0.835	0.426	0.873
0.60	0.793	0.414	0.890
0.70	0.793	0.452	0.915
0.80	0.762	0.451	0.940
0.90	0.463	0.290	0.950
0.95	0.207	0.226	0.944
1.00	0.049	0.204	1.000

[a]PPV: positive predictive value.

[b]NPV: negative predictive value.

Discussion

Principal Findings

The ability to predict a child's response to HFNC reliably and in real time could help guide clinical differentiation among three groups: (1) children most likely to do well on HFNC alone, (2) children most likely to need a higher level of support, and (3) children whose likely HFNC outcome is unclear and who may require additional observation. Patients identified from the first group may require less clinical intervention and free up scarce ICU resources. Identifying children in the second group may enable clinicians to intervene more rapidly and provide adequate support to prevent decompensation. Owing to clinical uncertainty, children in the third group may benefit from more careful and frequent observation with the continuous prediction of the likelihood of failure.

The granular longitudinal data captured from children in the ICU presents a tremendous amount of information available for learning and developing tools to help differentiate children's responses to numerous ICU interventions such as HFNC, including ventilation, extracorporeal membrane oxygenation, and dialysis. Deep learning models, especially those with sequential processing capabilities such as LSTMs, have the potential to use rich time-dependent data in ways that more traditional machine learning models (eg, LR) cannot; however, LSTMs may require sizable training data to construct generalizable models. The results from this study showed this

to be the case: a standard, 3-layer LSTM was generally the worst performing model on the hold-out test set.

TL was incorporated to address the issue of training LSTMs with insufficient data. The models with TL had the advantage of learning representations from >9000 PICU episodes, whereas the models without TL learned from >600 HFNC trials (approximately 500 episodes). The results demonstrate considerable gains from using TL and are consistent with the theory [14]. Figure 5, Figure S10, and Table S8, in particular, highlight the significant and time-independent performance increase delivered by TL in the LSTM models.

Input perseveration by itself (LSTM+3xPers) provided a performance boost relative to the standard LSTM, especially in the first 12 hours of the HFNC period in respiratory patients (Table S9, Multimedia Appendix 9). When combined with TL (LSTM+3xPers+TL), it continued to provide additional, although slight, gains. As demonstrated in the study by Ledbetter et al [18], LSTMs can exhibit a predictive lag phenomenon, wherein they fail to react rapidly to new data reflecting sudden clinical events and changes in patient status. In the context of HFNC use and decisions about whether to escalate a child to higher levels of support, this predictive lag may be deleterious in a time-constrained environment such as the PICU.

Finally, the ensemble models (Simple-EN-LSTM+3xPers+TL and Multi-EN-LSTM+3xPers+TL) were built to address another consequence of limited training data: the relatively high variability of any one particular realization of the model. This

XSL•FO

RenderX

is a byproduct of randomly chosen initialization seeds used to initialize LSTM weights and biases and for random dropout techniques used for regularization purposes. The ensemble methods provided a consistently higher performance on the hold-out test set than the nonensemble models. The ensemble models provided a slight performance boost over just a single LSTM+3xPers+TL model. Not surprisingly, the multiensembling of multiple models (both of those used to generate TL weights and those used to generate HFNC predictions) provided the best overall model (Multi-EN-LSTM+3xPers+TL).

Regardless of the model, the performance generally increased over time (Figure 5 and Figure S10). This is not surprising as the *lead time* (the interval between the times of prediction and outcome) decreases [25].

Model performance in patients with respiratory diagnoses is of interest as the pathophysiology of respiratory illness is particularly amenable to HFNC therapy [1-4]. Approximately 70% of HFNC initiation in this cohort were in patients with respiratory illnesses. Table S8 shows that all models except LR-14 generally performed better in the respiratory group over time. Figure 7 shows that the best performing models in the overall cohort—those that incorporated TL—performed even better in the respiratory group after 2 hours of observation, demonstrating the TL models' potential clinical impact.

The 2-hour mark after HFNC initiation is an important clinical decision point as a child has had adequate time to adapt to HFNC, and the effects of treatment can be assessed. This motivated the additional analyses of 2-hour predictions shown in Figure 6 and Figure 7 (ROC curves) and Table 2 (sensitivity, specificity, PPV, and NPV at various decision thresholds). The Multi-EN-LSTM+3xPers+TL model had the highest AUROC. In this model's ROC curve for the entire cohort, 2 operating points are of particular interest: the first corresponds to 95% sensitivity (20% specificity), and the second corresponds to 25% sensitivity (97% specificity, 67% PPV, and 84% NPV).

The first point can be used to identify children most likely to do well in HFNC (group 1), whereas the second can identify those most likely to fail HFNC (group 2). Successfully identifying 20% of group 1 can reduce the observational burden, whereas identifying 25% of group 2 could lead clinicians to intervene earlier with an escalation to a higher level of O_2 support, potentially improving outcomes for these children [9,10]. This system could potentially enable intervention 2 to 3 hours earlier in those most likely to fail HFNC. Children for whom the model predictions fall between the 2 thresholds are in the third group: those whose HFNC outcome is unclear and who may benefit from more frequent observations.

Limitations

This study had several limitations. First, it was based on a single-center retrospective cohort. Second, the target definition considered only the first 24 hours following HFNC initiation. Further work can refine the target to consider the subsequent 24 hours, regardless of how long the patient has already been on HFNC.

Finally, this study is limited by the exclusion of children experiencing apnea, making the predictive model's applicability to such children unclear. Although less than ideal, this exclusion was deemed necessary as it is difficult to determine whether escalation to BiPAP in these children is because of clinical necessity (ie, true escalation) or prophylactic caution (to guard against sleep apnea).

Conclusions

This study demonstrated the feasibility of applying advanced machine learning methodology to a complex and challenging clinical situation. This work demonstrated that clinically relevant models can be trained to predict the risk of escalation from HFNC within 24 hours of initiation of therapy and could be obtained by using an LSTM with the application of TL and input perseveration to boost AUROC performance.

Authors' Contributions

GP, DL, and MA were involved in the design, implementation, analysis, and writing of the manuscript. RW was involved in the design, analysis, and writing of the manuscript.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Demographic variables and vital observations used as input variables for long short-term memory models. See Table S5 for acronym expansions.
[DOCX File , 12 KB - medinform_v10i3e31760_app1.docx]

Multimedia Appendix 2
Laboratory and echocardiogram measurements used as input variables for long short-term memory models. See Table S5 for acronym expansions.

[DOCX File , 12 KB - medinform_v10i3e31760_app2.docx]

Multimedia Appendix 3

Drugs used as input variables for long short-term memory models. See Table S5 for acronym expansions.

[DOCX File , 12 KB - medinform_v10i3e31760_app3.docx]

Multimedia Appendix 4

Interventions used as input variables for long short-term memory models. See Table S5 for acronym expansions.

[DOCX File , 11 KB - medinform_v10i3e31760_app4.docx]

Multimedia Appendix 5

Acronyms and abbreviations used in Tables S1 to S4.

[DOCX File , 14 KB - medinform_v10i3e31760_app5.docx]

Multimedia Appendix 6

List of the input variables used in the 14-variable logistic regression model.

[DOCX File , 9 KB - medinform_v10i3e31760_app6.docx]

Multimedia Appendix 7

Number of high flow nasal cannula trials remaining at various evaluation points.

[DOCX File , 16 KB - medinform_v10i3e31760_app7.docx]

Multimedia Appendix 8

Test set area under the receiver operating characteristics (AUROCs) curve of the 8 models considered at various time points following high flow nasal cannula initiation. The highest AUROC in a column is in bold font.

[DOCX File , 10 KB - medinform_v10i3e31760_app8.docx]

Multimedia Appendix 9

Test set area under the receiver operating characteristics (AUROCs) curve of the 8 models considered at various time points following high flow nasal cannula initiation for children with a respiratory diagnosis. The highest AUROC in a column is in bold font.

[DOCX File , 10 KB - medinform_v10i3e31760_app9.docx]

Multimedia Appendix 10

Area under the receiver operating characteristics (AUROCs) curve of model predictions over the first 15 hours for a fixed set of patients with at least 15 hours of data on high flow nasal cannula on the holdout test set.

[PNG File , 234 KB - medinform_v10i3e31760_app10.png]

Multimedia Appendix 11

Area under the precision recall curves (AUPRCs) of model predictions over the first 15 hours for a fixed set of patients with at least 15 hours of data on high flow nasal cannula on the holdout test set.

[PNG File , 186 KB - medinform_v10i3e31760_app11.png]

Multimedia Appendix 12

Sensitivity versus specificity of the 2-hour predictions in the entire test set. The highest specificity along each row (fixed sensitivity) is in bold.

[DOCX File , 11 KB - medinform_v10i3e31760_app12.docx]

Multimedia Appendix 13

Sensitivity versus positive predictive value of the 2-hour predictions in the entire test set. The highest positive predictive value along each row (fixed sensitivity) is in bold.

[DOCX File , 10 KB - medinform_v10i3e31760_app13.docx]

Multimedia Appendix 14

Sensitivity versus negative predictive value of the 2-hour predictions in the entire test.

[DOCX File , 10 KB - medinform_v10i3e31760_app14.docx]

References

1. Kawaguchi A, Garros D, Joffe A, DeCaen A, Thomas NJ, Schibler A, et al. Variation in practice related to the use of high flow nasal cannula in critically ill children. Pediatr Crit Care Med 2020;21(5):228-235. [doi: 10.1097/pcc.0000000000002258]

2. Habra B, Janahi IA, Dauleh H, Chandra P, Veten A. A comparison between high-flow nasal cannula and noninvasive ventilation in the management of infants and young children with acute bronchiolitis in the PICU. Pediatr Pulmonol 2020 Feb 10;55(2):455-461. [doi: 10.1002/ppul.24553] [Medline: 31922360]

3. Clayton JA, McKee B, Slain KN, Rotta AT, Shein SL. Outcomes of children with bronchiolitis treated with high-flow nasal cannula or noninvasive positive pressure ventilation*. Pediatr Crit Care Med 2019;20(2):128-135. [doi: 10.1097/pcc.0000000000001798]

4. Ramnarayan P, Schibler A. Glass half empty or half full? The story of high-flow nasal cannula therapy in critically ill children. Intens Care Med 2017 Feb 26;43(2):246-249. [doi: 10.1007/s00134-016-4663-2] [Medline: 28124737]

5. Coletti KD, Bagdure DN, Walker LK, Remy KE, Custer JW. High-flow nasal cannula utilization in pediatric critical care. Respir Care 2017 Aug 06;62(8):1023-1029 [FREE Full text] [doi: 10.4187/respcare.05153] [Medline: 28588119]

6. Mikalsen I, Davis P, Øymar K. High flow nasal cannula in children: a literature review. Scand J Trauma Resusc Emerg Med 2016 Jul 12;24:93 [FREE Full text] [doi: 10.1186/s13049-016-0278-4] [Medline: 27405336]

7. Slain KN, Shein SL, Rotta AT. The use of high-flow nasal cannula in the pediatric emergency department. J Pediatr (Rio J) 2017 Nov;93 Suppl 1:36-45 [FREE Full text] [doi: 10.1016/j.jped.2017.06.006] [Medline: 28818509]

8. Kelly G, Simon H, Sturm J. High-flow nasal cannula use in children with respiratory distress in the emergency department. Pediatr Emerg Care 2013;29(8):888-892. [doi: 10.1097/pec.0b013e31829e7f2f]

9. Bauer P, Gajic O, Nanchal R, Kashyap R, Martin-Loeches I, Sakr Y, ICON Investigators (Supplemental Appendix 1). Association between timing of intubation and outcome in critically ill patients: a secondary analysis of the ICON audit. J Crit Care 2017 Dec;42:1-5. [doi: 10.1016/j.jcrc.2017.06.010] [Medline: 28641231]

10. Kang BJ, Koh Y, Lim C, Huh JW, Baek S, Han M, et al. Failure of high-flow nasal cannula therapy may delay intubation and increase mortality. Intens Care Med 2015 Apr 18;41(4):623-632. [doi: 10.1007/s00134-015-3693-5] [Medline: 25691263]

11. Guillot C, Le Reun C, Behal H, Labreuche J, Recher M, Duhamel A, et al. First-line treatment using high-flow nasal cannula for children with severe bronchiolitis: applicability and risk factors for failure. Arch Pediatr 2018 Apr;25(3):213-218. [doi: 10.1016/j.arcped.2018.01.003] [Medline: 29551475]

12. Er A, Çağlar A, Akgül F, Ulusoy E, Çitlenbik H, Yılmaz D, et al. Early predictors of unresponsiveness to high-flow nasal cannula therapy in a pediatric emergency department. Pediatr Pulmonol 2018 Jun 12;53(6):809-815. [doi: 10.1002/ppul.23981] [Medline: 29528202]

13. Yurtseven A, Saz E. The effectiveness of heated humidified high-flow nasal cannula in children with severe bacterial pneumonia in the emergency department. J Pediatr Res 2020 Mar 1;7(1):71-76. [doi: 10.4274/jpr.galenos.2019.15045]

14. Virtual Pediatric Systems (VPS). URL: https://www.myvps.org [accessed 2022-02-09]

15. Ho LV, Ledbetter D, Aczon M, Wetzel R. The dependence of machine learning on electronic medical record quality. AMIA Annu Symp Proc 2017;2017:883-891 [FREE Full text] [Medline: 29854155]

16. Imholz BP, Settels JJ, van der Meiracker AH, Wesseling KH, Wieling W. Non-invasive continuous finger blood pressure measurement during orthostatic stress compared to intra-arterial pressure. Cardiovasc Res 1990 Mar 01;24(3):214-221. [doi: 10.1093/cvr/24.3.214] [Medline: 2346955]

17. Schulman CS, Staul L. Standards for frequency of measurement and documentation of vital signs and physical assessments. Crit Care Nurse 2010 Jun;30(3):74-76 [FREE Full text] [doi: 10.4037/ccn2010406] [Medline: 20515885]

18. Ledbetter DR, Laksana E, Aczon M, Wetzel R. Improving recurrent neural network responsiveness to acute clinical events. IEEE Access 2021;9:106140-106151 [FREE Full text] [doi: 10.1109/access.2021.3099996]

19. Tan C, Sun F, Kong T, Zhang W, Yang C, Liu C. A survey on deep transfer learning. arXiv 2018;18(08) [FREE Full text]

20. Silver DL, Bennett KP. Guest editor's introduction: special issue on inductive transfer learning. Mach Learn 2008 Oct 21;73(3):215-220. [doi: 10.1007/s10994-008-5087-1]

21. Weiss K, Khoshgoftaar TM, Wang D. A survey of transfer learning. J Big Data 2016 May 28;3(1) [FREE Full text] [doi: 10.1186/s40537-016-0043-6]

22. Niu S, Liu Y, Wang J, Song H. A decade survey of transfer learning (2010–2020). IEEE Trans Artif Intell 2020 Oct;1(2):151-166. [doi: 10.1109/tai.2021.3054609]

23. Aczon M, Ledbetter D, Laksana E, Ho L, Wetzel R. Continuous prediction of mortality in the PICU: a recurrent neural network model in a single-center dataset. Pediatr Crit Care Med 2021 Jun 01;22(6):519-529 [FREE Full text] [doi: 10.1097/PCC.0000000000002682] [Medline: 33710076]

24. Ganaie M, Hu M, Tanveer M, Suganthan PN. Ensemble deep learning: a review. arXiv 2021;21(04) [FREE Full text]

25. Leisman DE, Harhay MO, Lederer DJ, Abramson M, Adjei AA, Bakker J, et al. Development and reporting of prediction models. Crit Care Med 2020;48(5):623-633. [doi: 10.1097/ccm.0000000000004246]

Abbreviations

AUROC: area under the receiver operating characteristic

BiPAP: bilevel positive airway pressure
CHLA: Children's Hospital Los Angeles
EMR: electronic medical record
HFNC: high flow nasal cannula
ICU: intensive care unit
LR: logistic regression
LR-14: 14-variable logistic regression
LR-517: 517-variable logistic regression
LSTM: long short-term memory
ML: machine learning
NIV: noninvasive ventilation
NPV: negative predictive value
PICU: pediatric intensive care unit
PPV: positive predictive value
ROC: receiver operating characteristic
TL: transfer learning

Original Paper

Predicting Long-term Survival After Allogeneic Hematopoietic Cell Transplantation in Patients With Hematologic Malignancies: Machine Learning–Based Model Development and Validation

Eun-Ji Choi[1*], MD, PhD; Tae Joon Jun[2*], PhD; Han-Seung Park[1], MD; Jung-Hee Lee[1], MD, PhD; Kyoo-Hyung Lee[1], MD, PhD; Young-Hak Kim[3], MD, PhD; Young-Shin Lee[1], MSN, PhD; Young-Ah Kang[1], MSN; Mijin Jeon[1], MSN; Hyeran Kang[1], MSN; Jimin Woo[1], MSN; Je-Hwan Lee[1], MD, PhD

[1]Department of Hematology, Asan Medical Center, University of Ulsan College of Medicine, Seoul, Republic of Korea

[2]Big Data Research Center, Asan Institute for Life Sciences, Asan Medical Center, Seoul, Republic of Korea

[3]Division of Cardiology, Asan Medical Center, University of Ulsan College of Medicine, Seoul, Republic of Korea

[*]these authors contributed equally

Corresponding Author:
Je-Hwan Lee, MD, PhD
Department of Hematology
Asan Medical Center
University of Ulsan College of Medicine
88, Olympic-ro 43-gil
Songpa-gu
Seoul, 05505
Republic of Korea
Phone: 82 3010 3218
Fax: 82 3010 6961
Email: jhlee3@amc.seoul.kr

Abstract

Background: Scoring systems developed for predicting survival after allogeneic hematopoietic cell transplantation (HCT) show suboptimal prediction power, and various factors affect posttransplantation outcomes.

Objective: A prediction model using a machine learning–based algorithm can be an alternative for concurrently applying multiple variables and can reduce potential biases. In this regard, the aim of this study is to establish and validate a machine learning–based predictive model for survival after allogeneic HCT in patients with hematologic malignancies.

Methods: Data from 1470 patients with hematologic malignancies who underwent allogeneic HCT between December 1993 and June 2020 at Asan Medical Center, Seoul, South Korea, were retrospectively analyzed. Using the gradient boosting machine algorithm, we evaluated a model predicting the 5-year posttransplantation survival through 10-fold cross-validation.

Results: The prediction model showed good performance with a mean area under the receiver operating characteristic curve of 0.788 (SD 0.03). Furthermore, we developed a risk score predicting probabilities of posttransplantation survival in 294 randomly selected patients, and an agreement between the estimated predicted and observed risks of overall death, nonrelapse mortality, and relapse incidence was observed according to the risk score. Additionally, the calculated score demonstrated the possibility of predicting survival according to the different transplantation-related factors, with the visualization of the importance of each variable.

Conclusions: We developed a machine learning–based model for predicting long-term survival after allogeneic HCT in patients with hematologic malignancies. Our model provides a method for making decisions regarding patient and donor candidates or selecting transplantation-related resources, such as conditioning regimens.

XSL·FO
RenderX

Introduction

Background

Allogeneic hematopoietic cell transplantation (HCT) is a potentially curative therapeutic option for patients with hematologic malignancies, which has been widely used. The increasing use of allogeneic HCT is attributable to multiple factors, including improved alternative donor availability, reduced-intensity conditioning regimens, advances in the prevention of transplantation-related toxicities, and an improvement in general supportive care. Despite these advances, allogeneic HCT remains associated with considerably high rates of complications, treatment-related mortality, and relapse [1]. To predict transplantation outcomes more accurately before making decisions regarding transplant eligibility, several prognostic scoring systems for survival after allogeneic HCT have been developed. These scores include the HCT-specific comorbidity index, the European Group for Blood and Marrow Transplantation (EBMT) risk score, and the Pretransplant Assessment of Mortality score, among others [2-5]. Most prognostic scoring systems were developed using parametric statistical methodologies, such as Cox proportional hazards models, for predicting the likelihood of survival for HCT recipients. The scoring systems' variables mainly include recipient factors, such as age, comorbidities, performance status, time from diagnosis to HCT, and disease status. Furthermore, several donor factors are considered in the EBMT risk score, including donor type (human leukocyte antigen [HLA]: identical sibling or matched unrelated), sex match, and cytomegalovirus serostatus. However, the reported accuracy of these prediction models is suboptimal, where the area under the receiver operating characteristic (ROC) curve (AUC) ranges between 0.6 and 0.7 [6].

Recently, attempts to predict transplantation-related outcomes more accurately have been made in various clinical settings regarding early mortality [7], graft-versus-host disease (GVHD) [8], or relapse using deep learning–based prediction models [9]. The Acute Leukemia (AL)–EBMT score was developed using a data mining–based approach to predict 100-day mortality after allogeneic HCT [7]. Another study of a machine learning algorithm predicting the incidence of acute GVHD using Japanese registry data has demonstrated that the calculated scores were associated with clear stratification of acute GVHD, whereas lower scores were associated with a low incidence of acute GVHD [8]. In one study, a machine learning–based model was developed to predict the 1-year relapse rate after allogeneic HCT in patients with AL [9]. Although the patient population, endpoints, and criteria for variable selection were different between studies, the performances were similar and seemed slightly better than the previously reported transplantation outcome prediction scores.

The survival following allogeneic HCT, however, can vary depending on multiple variables, such as disease relapse and transplantation-related complications, including GVHD, engraftment failure, or infection, which can lead to increased nonrelapse mortality (NRM). Furthermore, these HCT complications are associated with several variables, including donor-related or recipient-related factors, donor-recipient relationship, and conditioning, among others.

Objective

We hypothesized that the selection of variables using a machine learning–based approach and the establishment of a prediction model by applying those variables will improve the performance of the model and avoid unexpected biases. Additionally, we assumed that the established prediction algorithm will help choose better transplantation-related factors or donors to improve post-HCT outcomes. In this study, we developed a model for predicting the long-term survival of patients with hematologic malignancies after allogeneic HCT based on selected variables using a machine learning algorithm, and we validated the model's accuracy in a validation set. Then, we implemented an algorithm to select more appropriate transplantation-related factors using the established prediction model.

Methods

Patient Population and Study Outcomes

Data on 1470 adult patients (≥15 years old) with hematologic malignancies who underwent allogeneic HCT between December 1993 and December 2015 at Asan Medical Center, Seoul, South Korea, were obtained for developing the machine learning–based prediction model. To predict long-term survival after allogeneic HCT, we included patients who survived more than 5 years and who died within 5 years after transplantation. As the data cutoff date was December 2020, we only included patients who underwent allogeneic HCT before December 2015 to ensure that the follow-up duration of each patient could be at least 5 years. Then, 229 variables, including recipient and donor characteristics, disease features, HLA types, graft information, administered medications for conditioning, GVHD prophylaxis, supportive care, and other laboratory data, were collected for analysis.

The primary objective of the study was to predict the 5-year overall survival (OS) after allogeneic HCT, and the secondary objectives include determining the NRM, cumulative incidence of relapse (CIR), and 100-day OS. All censored data were calculated from the date of the transplantation.

Ethics Approval

The Institutional Review Board of Asan Medical Center approved the protocols of this study (2021-1003), which was conducted according to the 2008 Declaration of Helsinki.

Selection of a Predicting Model

The patients were classified into two groups, those who survived more than 5 years and those who died within 5 years. In the learning process, the former group was labeled 0 and the latter was labeled 1. Therefore, the closer the predicted value to 1, the higher the probability of death within 5 years. The aforementioned predictive factors were classified into categorical or noncategorical variables and used for developing 5 prediction models. The performance for predicting survival after allogeneic HCT was tested using the following 5 machine learning algorithms: gradient boosting machine (GBM), random

forest, deep neural network, logistic regression, and adaptive boosting (AdaBoost). Each algorithm was tested using the same training set which was randomly divided (1176/1470, 79.59% of the total number of patients in the training set). The AUCs of the algorithms are shown in Multimedia Appendix 1. Of the 5 algorithms, GBM showed the highest AUC (0.75) compared with random forest (0.74), deep neural network (0.65), logistic regression (0.70), and AdaBoost (0.72). Therefore, the selection of relevant variables and the development of the final model were performed using GBM. GBM is an ensemble method that combines several weak classifiers, such as trees. The goal of GBM is to focus and place the weights on incorrectly predicted results through gradient descent [10]. While GBM is training, the initial tree trains the data set and assigns weights to incorrectly predicted records with errors, and the next tree from the same model learns the weighted data set and repeats the process of assigning weights.

Explainable Individualized Survival Prediction

We provided an explainable individualized survival prediction using Shapley values to quantify the probability of surviving for each patient by predicting the OS after allogeneic HCT. A Shapley value is calculated as the average change according to the presence or absence of a single feature over all possible combinations of features [11]. Given a survival prediction model, f(x), we can compute the Shapley values using the following equation:

where n is the total number of features, and the sum extends over all subsets S of N not containing feature i. In a recent study, a unified framework called Shapley Additive Explanations (SHAP) was released for explainable machine learning models using Shapley values [10]. In this study, the survival model also provides a description of patient-specific survival prediction using SHAP.

Other Statistical Analyses

Categorical variables were compared using the chi-square test or Fisher exact test, and continuous variables were compared using the Mann-Whitney U-test or Student t test, as appropriate. The OS was calculated using the Kaplan-Meier method, and the resulting survival curves were compared using the log-rank test. NRM and CIR were evaluated using a cumulative incidence function regarding competing risks and compared using the method of Gray in R, version 3.6.3 (R Foundation for Statistical Computing). All statistical analyses were conducted using SPSS, version 24 (IBM Corp), and graphs were generated using GraphPad Prism, version 9.1.2 (GraphPad Software Inc). In all analyses, P values were two-tailed, and those less than .05 were used to denote statistical significance.

Results

Patient and Donor Characteristics

The characteristics of the patients and donors included in the study are shown in Table 1. Between December 2009 and December 2015, 1470 patients underwent allogeneic HCT for hematologic malignancies, including acute myeloid leukemia (n=783), acute lymphoblastic leukemia (ALL; n=306), myelodysplastic syndrome (MDS; n=188), chronic myeloid leukemia (n=92), non-Hodgkin/Hodgkin lymphoma (n=56), BCR-ABL1–negative myeloproliferative neoplasm (MPN; n=16), MDS/MPN (n=6), and multiple myeloma (n=13). Approximately two-thirds of the patients (n=995) received peripheral blood as a graft source, and one patient who received cord blood as a graft source was included. Reduced-intensity conditioning and myeloablative conditioning were used in 934 (63.5%) and 536 (36.5%) of the 1470 patients, respectively. Antithymocyte globulin was used in 903 (61.4%) of the 1470 patients as GVHD prophylaxis.

During the median follow-up duration of 8 years (95% CI 7.8-8.3 years), the estimated 5-year OS of all patients was 46.2%. The 2-year incidence of NRM and CIR was 17.7% and 33.3%, respectively.

XSL·FO

RenderX

Table 1. Patient and donor characteristics.

Variable	Value
Patients, N	1470
Interval between diagnosis to HCT[a] in months, median (95% CI)	5.7 (0-268)
Recipient sex, n (%)	
Male	833 (56.7)
Female	637 (43.3)
Donor sex, n (%)	
Male	977 (66.5)
Female	493 (33.5)
Recipient age in years, median (range)	41 (15-75)
Donor age in years, median (range)	34 (0-70)
Donor-recipient sex, n (%)	
Male to male	551 (37.5)
Female to male	280 (19)
Male to female	424 (28.8)
Female to female	213 (14.5)
Recipient disease, n (%)	
AML[b]	783 (66.9)
MDS[c]	188 (16.1)
ALL[d]	306 (26.2)
Lymphoma	56 (4.8)
MM[e]	13 (1.1)
CML[f]	92 (7.9)
MPN[g]	16 (1.4)
MDS-MPN	16 (1.4)
HCT-CI[h] score, median (range)	3 (0-8)
Disease risk, n (%)	
Standard risk[i]	830 (56.5)
High risk	640 (43.5)
Donor type, n (%)	
Matched sibling	591 (40.2)
Unrelated	387 (26.4)
Haploidentical familial	491 (33.4)
Cord blood	1 (0.1)
Graft source, n (%)	
Bone marrow	472 (32.1)
Peripheral blood	997 (67.8)
Cord blood	1 (0.1)
Conditioning intensity, n (%)	
Myeloablative	536 (36.5)
Reduced intensity	934 (63.5)

Variable	Value
Treated with antithymocyte globulin to prevent GVHD[j], n (%)	903 (61.4)

[a]HCT: hematopoietic cell transplantation.

[b]AML: acute myeloid leukemia.

[c]MDS: myelodysplastic syndrome.

[d]ALL: acute lymphoblastic leukemia.

[e]MM: multiple myeloma.

[f]CML: chronic myeloid leukemia.

[g]MPN: myeloproliferative neoplasm.

[h]HCT-CI: hematopoietic cell transplantation–specific comorbidity index.

[i]The standard-risk group is defined as follows: patients with acute leukemia in the first remission (except by salvage chemotherapy), CML in the chronic phase, drug-sensitive lymphoma/MM, or MDS with bone marrow blasts ≤5% at HCT.

[j]GVHD: graft-versus-host disease.

Development of the Prediction Model

After deciding on GBM as the prediction algorithm, the variables used for model development were selected using the recursive feature elimination (RFE) method. RFE is one of the widely used feature selection methods that provide a rank to each variable according to feature importance in predicting the target variable and help select a minimum specified number of variables showing good performance in a model [12]. Using the RFE algorithm, we selected 45 relevant variables for developing the prediction model (see Multimedia Appendix 2 and Textbox 1). In the case of HLA type, each allele of recipients and donors was regarded as an independent variable.

Textbox 1. Selected variables for the prediction model. AML: acute myeloid leukemia. WBC: white blood cell. HLA: human leukocyte antigen. RBC: red blood cell. CMV: cytomegalovirus. HCT-CI: hematopoietic cell transplantation–specific comorbidity index. *The standard-risk group is defined as follows: patients with acute leukemia in the first remission (except by salvage chemotherapy), CML in the chronic phase, drug-sensitive lymphoma/MM, or MDS with bone marrow blasts ≤5% at HCT.


Variables

- Diagnosis and disease (eg, AML first complete remission)
- Disease risk*
- WBC count at diagnosis
- Extramedullary disease at diagnosis
- Extramedullary disease at HCT
- Karyotype at diagnosis
- Karyotype at HCT
- CMV serostatus of recipient
- CMV serostatus of donor
- Hepatic score of HCT-CI
- Total score of HCT-CI
- Conditioning regimen
- Donor type
- Recipient HLA type: A, B, C, DR, and DQ
- Donor HLA type: A, B, C, DR, and DQ
- RBC transfusion before HCT
- Platelet transfusion before HCT

Final Performance of the Prediction Model

The performance of the prediction model using GBM and selected variables in 294 patients is depicted in Figure 1A. The AUC and prediction accuracy of the final model were 0.788 and 0.712, respectively. Figure 1B shows a calibration plot with the Brier score of the model demonstrating agreement between the estimated predicted risk and observed risk of death in the validation cohort. The algorithm was trained and evaluated using 10-fold cross-validation in the total patient cohort, where the predictive power of the model demonstrated a generalized performance with a similar accuracy.

Figure 1. The final performance of the prediction model. Panel A shows the area under the receiver operating characteristic curve. Panel B shows the calibration plot.
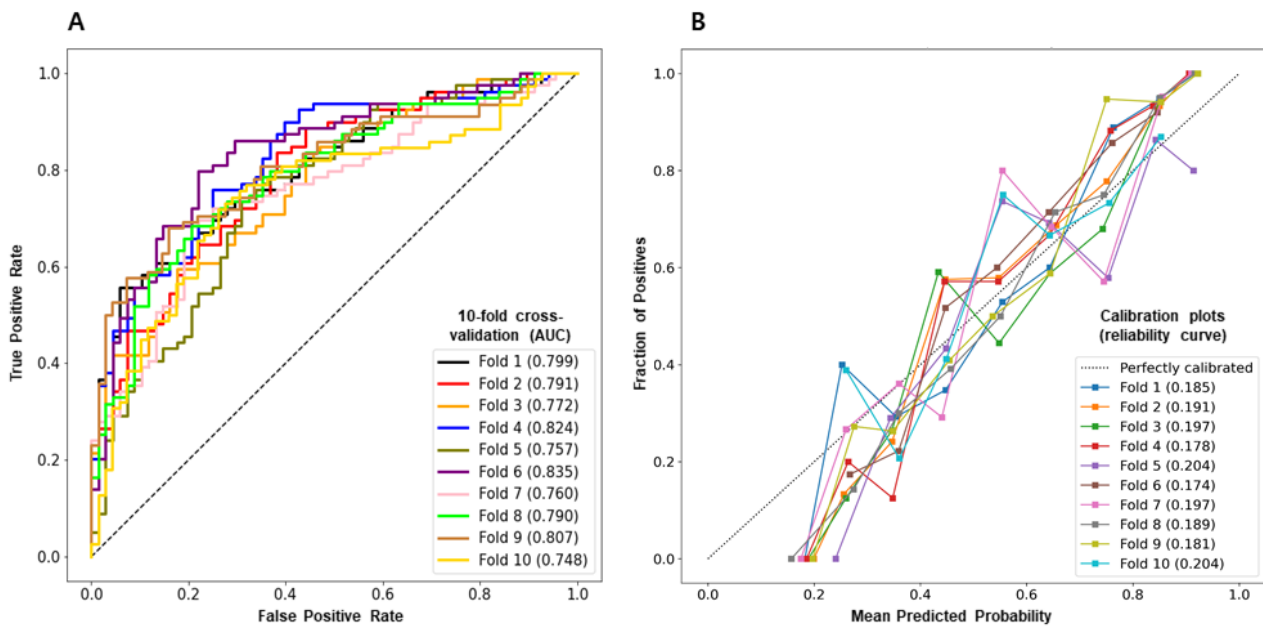
Because we classified patients who died within 5 years as 1, the closer the predicted value of the GBM model to 1, the higher the risk of death. The optimal threshold for determining whether the risk score is positive or negative is calculated using the Youden J statistic along with the ROC curve. From the prediction model, the threshold is 0.5533, and if the risk score is greater than that, the model estimates that the patient will die within 5 years. The predicting probability of the risk score of

each patient was tested in a randomly selected patient cohort, which corresponds to 20% of all patients (294/1470) to reduce the probable bias from choosing one of 10 produced models. Figure 2A shows the estimated post-transplantation OS of the patients according to the risk score, which was equally divided by the absolute score values. The estimated 5-year OS was 70.3% in the low-risk group, 42.6% in the intermediate-risk group, and 14.9% in the high-risk group ($P<.001$).

Figure 2. Different post-transplantation outcomes of the patients of validation set according to the prediction score (A) Overall survival (B) relapse (C) non-relapse mortality.
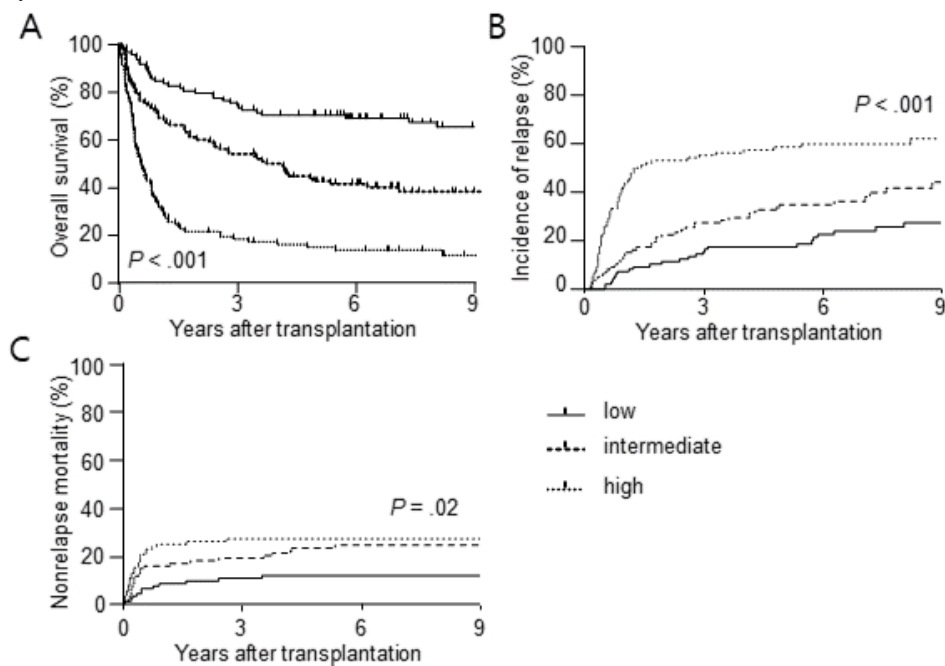
Prediction of NRM and Relapse

To assess whether the risk score can also predict NRM and relapse after HCT, we analyzed the incidence of NRM and relapse using 3 risk groups. High-risk scores were significantly

associated with both higher CIR ($P<.001$) and higher NRM ($P=.02$) (Figure 2B and 2C). The estimated 2-year CIR was 11.3% in the low-risk group, 22.4% in the intermediate-risk group, and 53.1% in the high-risk group. The 2-year NRM was

9.3% in the low-risk group, 17.3% in the intermediate-risk group, and 25% in the high-risk group.

Application of the Algorithm for Donor Selection

We assumed that the prediction score for each patient can be applied in selecting the most appropriate donor when there are multiple donor candidates. For example, the prediction score can help physicians select the donor between a younger HLA-haploidentical individual and an older matched sibling. To verify this, we calculated the scores using Shapley values through which the importance of each variable can be visualized using a specific value. We simulated a real case of a patient with ALL in the first CR who has the following 2 donor candidates: one is a 48-year-old HLA-haploidentical familial female individual, and the other is a 43-year-old locus-mismatched unrelated male individual. A total of 2 prediction scores were calculated using data derived from each donor showing different values (Figure 3). Among the variables, the pretransplantation disease status appeared to be the most important factor in calculating each score. According to our prediction model, the donor in Figure 3B (unrelated donor) could be preferred because the score is lower than the donor in Figure 3A (haploidentical donor).

Figure 3. Survival difference of the patients of validation set according to the prediction score.
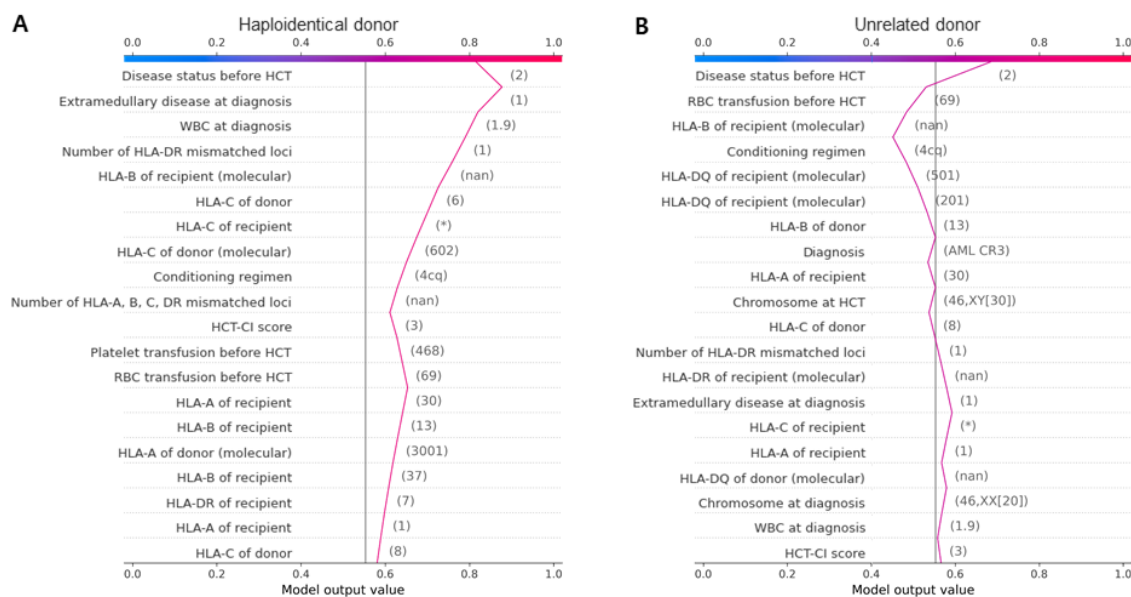
Discussion

Principal Findings

Long-term survival after allogeneic HCT in patients with hematologic malignancies is affected by multiple factors but mainly depends on disease relapse and NRM. Multiple variables, including disease status, genetic risk, conditioning regimen, comorbidities, degree of HLA matching, and patient and donor ages, are associated with disease relapse, GVHD, engraftment, or treatment-related toxicities, and these outcomes are closely and mutually related to survival after transplantation. However, traditional statistical methods are unsuitable for analysis considering the interactions between variables or their differences according to the specific values of each factor, such as the relationship between the HLA allele of the patient and donor. In this regard, prediction models based on machine learning algorithms can be an effective alternative for predicting posttransplantation outcomes and can provide guidance for selecting appropriate patients, donors, or resources [13].

We developed a prediction model and risk score using GBM and selected variables based on machine learning for long-term survival after allogeneic HCT. Our model demonstrated an AUC of 0.788, which showed better performance in predicting posttransplantation outcomes than previously reported machine learning–based models. Shouval et al [7] have reported the

AL-EBMT model predicting 100-day mortality after allogeneic HCT showing an AUC of 0.701, which was significantly better than that of the EBMT score (AUC, 0.646). A study on Japanese individuals who underwent HCT has developed a machine learning–based prediction algorithm of acute GVHD, and the AUC of the model was 0.62 [8]. Another prediction algorithm developed by Fuse et al [9] has shown an AUC of 0.667 for predicting relapse within 1 year after transplantation. Most models were developed by applying the alternating decision tree algorithm, and the variables were selected by researchers. In this study, the model was developed using variables derived from the GBM algorithm and using the RFE method, instead of using preselected variables based on the opinion of the researchers or conventional statistical analysis. Through RFE, we extracted the minimum required features where the performance of the predicting model does not deteriorate. This is an important difference from the existing literature that applied machine learning algorithms using clinically selected variables. In contrast, we first built a full model using all possible variables and then gradually removed features that had little effect on survival prediction. Those differences might contribute to the higher AUC of our prediction model by reducing biases in selecting variables and augmenting possible correlations between each factor. Interestingly, the selected variables for our prediction model include each HLA allele type of recipients and donors. Because we used the raw values of

each HLA allele of both recipients and donors rather than calculating the degree of mismatch, direction of mismatch, or allele types, our approach integrated the interactions between alleles affecting survival.

To apply the prediction model to patients planning for allogeneic HCT in practice, a specific tool for comparing the expected outcomes according to multiple different factors is required. We provided a prediction score to quantify the probability of survival, which showed good concordance of the observed and estimated survival after HCT. Additionally, SHAP visualizes the importance of each factor (Figure 3), which allows for the prioritization of more appropriate transplantation-related resources. The most remarkable aspect of our model is that the importance of each factor can be quantified and visualized so that physicians can use the algorithm when planning allogeneic HCT to select factors, such as donor or conditioning regimen, that are expected to achieve better survival.

The limitations of this study include the relatively small number of patients used for establishing the algorithm-based prediction model. Although the model showed consistency using 10-fold cross-validation in the validation cohort, a larger patient cohort is considered more helpful in verifying the performance of the algorithm. Further external validation using data from a greater number of patients is warranted. Second, the retrospective nature of the study may have resulted in selection and measurement biases. However, we included all patients with hematologic malignancies who underwent allogeneic HCT during a certain period of time to reflect real-world practice.

Conclusions

Here, we present a machine learning–based algorithm and prediction score for quantifying the probability of long-term survival after allogeneic HCT in patients with hematologic malignancies. The prediction score showed a moderate negative correlation with long-term survival, NRM, and relapse after transplantation. Our prediction model provides a personalized method for selecting more appropriate transplantation-related factors and patient or donor candidates for allogeneic HCT.

Conflicts of Interest

None declared.

Multimedia Appendix 1
The AUC of each tested algorithm. cb, CatBoost; rf, random forest; fnn, feedforward neural network; log, logistic regression; ada, AdaBoost.
[PNG File , 92 KB - medinform_v10i3e32313_app1.png]

Multimedia Appendix 2
Recursive feature elimination showing the AUC (area under the curve) according to the number of selected features.
[PNG File , 45 KB - medinform_v10i3e32313_app2.png]

References

1. Phelan R, Arora M, Chen M. Current use and outcome of hematopoietic stem cell transplantation: CIBMTR US summary slides. Center for International Blood and Marrow Transplant Research. URL: https://www.cibmtr.org/ReferenceCenter/SlidesReports/SummarySlides/pages/index.aspx [accessed 2022-01-31]

2. Sorror ML, Maris MB, Storb R, Baron F, Sandmaier BM, Maloney DG, et al. Hematopoietic cell transplantation (HCT)-specific comorbidity index: a new tool for risk assessment before allogeneic HCT. Blood 2005 Oct 15;106(8):2912-2919 [FREE Full text] [doi: 10.1182/blood-2005-05-2004] [Medline: 15994282]

3. Gratwohl A, Stern M, Brand R, Apperley J, Baldomero H, de Witte T, et al. Risk score for outcome after allogeneic hematopoietic stem cell transplantation. Cancer 2009 Oct 15;115(20):4715-4726. [doi: 10.1002/cncr.24531]

4. Parimon T, Au DH, Martin PJ, Chien JW. A risk score for mortality after allogeneic hematopoietic cell transplantation. Ann Intern Med 2006 Mar 21;144(6):407. [doi: 10.7326/0003-4819-144-6-200603210-00007]

5. Armand P, Kim HT, Logan BR, Wang Z, Alyea EP, Kalaycio ME, et al. Validation and refinement of the Disease Risk Index for allogeneic stem cell transplantation. Blood 2014 Jul 05;123(23):3664-3671 [FREE Full text] [doi: 10.1182/blood-2014-01-552984] [Medline: 24744269]

6. Potdar R, Varadi G, Fein J, Labopin M, Nagler A, Shouval R. Prognostic scoring systems in allogeneic hematopoietic stem cell transplantation: where do we stand? Biol Blood Marrow Transplant 2017 Nov;23(11):1839-1846 [FREE Full text] [doi: 10.1016/j.bbmt.2017.07.028] [Medline: 28797781]

7. Shouval R, Labopin M, Bondi O, Mishan-Shamay H, Shimoni A, Ciceri F, et al. Prediction of allogeneic hematopoietic stem-cell transplantation mortality 100 days after transplantation using a machine learning algorithm: a European Group

for Blood and Marrow Transplantation Acute Leukemia Working Party retrospective data mining study. J Clin Oncol 2015 Oct 01;33(28):3144-3151. [doi: 10.1200/JCO.2014.59.1339] [Medline: 26240227]

8. Arai Y, Kondo T, Fuse K, Shibasaki Y, Masuko M, Sugita J, et al. Using a machine learning algorithm to predict acute graft-versus-host disease following allogeneic transplantation. Blood Adv 2019 Nov 26;3(22):3626-3634 [FREE Full text] [doi: 10.1182/bloodadvances.2019000934] [Medline: 31751471]

9. Fuse K, Uemura S, Tamura S, Suwabe T, Katagiri T, Tanaka T, et al. Patient-based prediction algorithm of relapse after allo-HSCT for acute leukemia and its usefulness in the decision-making process using a machine learning approach. Cancer Med 2019 Sep;8(11):5058-5067 [FREE Full text] [doi: 10.1002/cam4.2401] [Medline: 31305031]

10. Lundberg S, Lee S. A unified approach to interpreting model predictions. In: Proceedings of the 31st Annual Conference on Neural Information Processing Systems (NIPS). 2017 Presented at: 2017 Conference on Neural Information Processing Systems; December 4-9, 2017; Long Beach, CA.

11. Shapley L. A value for n-person games. In: Kuhn H, Tucker A, editors. Contributions to the Theory of Games II. Princeton, NJ: Princeton University Press; 1953:307-317.

12. Guyon I, Weston J, Barnhill S, Vapnik V. Gene selection for cancer classification using support vector machines. Mach learn 2002;46(1):389-422. [doi: 10.1023/a:1012487302797]

13. Shouval R, Bondi O, Mishan H, Shimoni A, Unger R, Nagler A. Application of machine learning algorithms for clinical predictive modeling: a data-mining approach in SCT. Bone Marrow Transplant 2013 Oct 7;49(3):332-337. [doi: 10.1038/bmt.2013.146]

Abbreviations

ALL: acute lymphoblastic leukemia
AUC: area under the ROC curve
CIR: cumulative incidence of relapse
GBM: gradient boosting machine
GVHD: graft-versus-host disease
HCT: hematopoietic cell transplantation
HLA: human leukocyte antigen
MDS: myelodysplastic syndrome
MPN: myeloproliferative neoplasm
NRM: nonrelapse mortality
OS: overall survival
RFE: recursive feature elimination
ROC: receiver operating characteristic
SHAP: Shapley Additive Explanations

Original Paper

Web-Based Skin Cancer Assessment and Classification Using Machine Learning and Mobile Computerized Adaptive Testing in a Rasch Model: Development Study

Ting-Ya Yang[1*], MD; Tsair-Wei Chien[2*], MBA; Feng-Jie Lai[3*], MD, PhD

[1]Department of Family Medicine, Chi Mei Medical Center, Tainan, Taiwan

[2]Department of Medical Research, Chi-Mei Medical Center, Tainan, Taiwan

[3]Department of Dermatology, Chi-Mei Medical Center, Tainan, Taiwan

[*]all authors contributed equally

Corresponding Author:
Feng-Jie Lai, MD, PhD
Department of Dermatology
Chi-Mei Medical Center
901, Zhonghua Rd
Yongkang District
Tainan, 710
Taiwan
Phone: 886 6 2812811 ext 57109
Fax: 886 6 2203706
Email: lai.fengjie@gmail.com

Abstract

Background: Web-based computerized adaptive testing (CAT) implementation of the skin cancer (SC) risk scale could substantially reduce participant burden without compromising measurement precision. However, the CAT of SC classification has not been reported in academics thus far.

Objective: We aim to build a CAT-based model using machine learning to develop an app for automatic classification of SC to help patients assess the risk at an early stage.

Methods: We extracted data from a population-based Australian cohort study of SC risk (N=43,794) using the Rasch simulation scheme. All 30 feature items were calibrated using the Rasch partial credit model. A total of 1000 cases following a normal distribution (mean 0, SD 1) based on the item and threshold difficulties were simulated using three techniques of machine learning—naïve Bayes, k-nearest neighbors, and logistic regression—to compare the model accuracy in training and testing data sets with a proportion of 70:30, where the former was used to predict the latter. We calculated the sensitivity, specificity, receiver operating characteristic curve (area under the curve [AUC]), and CIs along with the accuracy and precision across the proposed models for comparison. An app that classifies the SC risk of the respondent was developed.

Results: We observed that the 30-item k-nearest neighbors model yielded higher AUC values of 99% and 91% for the 700 training and 300 testing cases, respectively, than its 2 counterparts using the hold-out validation but had lower AUC values of 85% (95% CI 83%-87%) in the k-fold cross-validation and that an app that predicts SC classification for patients was successfully developed and demonstrated in this study.

Conclusions: The 30-item SC prediction model, combined with the Rasch web-based CAT, is recommended for classifying SC in patients. An app we developed to help patients self-assess SC risk at an early stage is required for application in the future.

XSL•FO
RenderX

Introduction

Background

Skin cancer (SC) is the most common malignant neoplasm occurring in White populations, and it is mainly divided into (1) malignant melanoma (MM) and (2) nonmelanoma SCs (NMSCs), which include squamous cell carcinoma and basal cell carcinoma as the major subtypes. The global incidence of MM and NMSCs is well-established and on the rise. In Australia, SC accounts for most newly diagnosed cancers each year, with age-standardized incidence rates for MM of 65.3×10^{-5} and 1878×10^{-5} for NMSCs [1,2]. There are >434,000 people in a population of only 23 million who treat keratinocyte cancer each year in Australia [2], causing a substantial socioeconomic burden and impact on public health services.

There are several well-recognized risk factors that increase the potential for the development of SC and have been reported in previous literature, such as UV radiation, genetic susceptibility, smoking, ionizing radiation, and the use of photosensitizing drugs [3]. Among the aforementioned risk factors, excessive UV radiation exposure remains the major causative risk factor for SC [4]. Therefore, it is crucial to modify personal behaviors to reduce direct and excessive sun exposure, such as avoiding long-term sunbathing or the use of indoor tanning devices, appropriately applying sunscreens, using sun-protective cloth garments, and staying in the shade.

Requirement for Prediction Model in Classification of SC

In practice, it is difficult to provide people with their individual risk of SC [1]. Given the lack of clear recommendations for organized SC screening, physical exploration, clinical history of lesion changes, and correlated family SC history continue to be key for detecting skin neoplasms. Assuming that a person has attributes that highly correlate with the underlying architecture of the skin, the potential risk of SC can be assessed through questions (ie, questionnaire items); for example, underlying pigmentation traits include hair color, eye color, the propensity to freckle and sunburn, skin phenotypes, and some personal behavior factors such as tanning attitudes and sunbed use. Accordingly, it is feasible to construct a unidimensional scale to measure these attributes using the responses to the unidimensional items and further calculate an overall SC risk score using an assessment tool (eg, web-based computerized adaptive testing [CAT] administrations [1]) or even classify the SC risk for patients in clinical settings.

Predicting SC Risk and Classifying the SC Possibility

Statistical validity is based on the correlations among item measures (or scores) on a questionnaire and people's unobservable true status (eg, melanoma status–deemed latent traits that cannot be directly detectable in the real world) [5]. The Rasch model [6] is a mathematical modeling approach that has been used to assess how well the items measure the underlying latent traits [7-13], which are based on a unidimensional scale when the data fit the Rasch model's expectation (ie, all items can be added to a summation score) [10,13]. Nonetheless, no SC classifications that use machine

learning to predict SC risk have been illustrated and demonstrated in the literature. We are motivated to develop a prediction model for classifying SC in adults who are potentially at risk.

CAT Assessment and Limitation in SC Classification

CAT is a tailored measure based on item response theory (IRT) [14,15] that can better align with each examinee's ability level [10,13,16]. The computer follows an IRT-based algorithm, and the difficulty of the next selected item depends mainly on all previously answered items. As such, each patient needs to answer the fewest possible items by dynamically selecting appropriate testing items, resulting in less respondent burden without compromising measurement precision and thereby making it possible to individualize each participant's assessment [1,10,13].

The limitation of CAT applied to machine learning is the missing responses (ie, unanswered items) in the data. Fortunately, generating the expected responses to endorse the answers in CAT has been resolved to overcome the drawback of not having all the items answered in CAT (ie, using the expected value to fill in the missing data, as done in previous studies [13,17,18]). As such, convolutional neural networks (CNNs) [19,20] combined with the expected responses to classify the groups of individual bullying levels [13] are applicable. Thus, we are interested in applying the expected responses to CAT to (1) reduce participant burden with more accurate outcomes [1,10,13,16] and (2) predict SC classification in patients.

Web-Based Assessment Using Smartphones

With the advent of the era of digital technology, the advancement and maturation of mobile health and health communication technology have been rapidly increasing [21]. To date, smartphone apps for classifying SC using CAT-based machine learning for patients in health care settings are lacking when searching for publications in the PubMed library using the keywords *skin and cancer* AND *computerized adaptive testing* AND *CAT* AND *machine learning* as of December 5, 2021. It is not only the complexity of the CAT procedure with multimedia illustrations embedded into a web-based module but also the difficulty of the model's parameters that need to be transformed into the probability of classification types when SC is assessed on the web. A web-based CAT app incorporating machine learning and SC could provide patients with a better understanding of the SC classification and prediction of SC at risk before a serious SC problem occurs.

Study Aims

The aims of this study are to (1) compare the prediction accuracy of SC between machine learning models in SC classification and (2) build a CAT-based SC assessment using machine learning to develop an app for automatic classification of SC to help patients assess SC risk at an early stage.

Methods

Data Source

On the basis of a previous study [1,22], we extracted data from a population-based Australian cohort study of SC risk (N=43,794) by simulating Rasch data [23], including 1000 virtual patients across 30 feature variables defined in the previous study [1] (Multimedia Appendix 1).

All data used in this study were simulated and extracted from the previous article [1]. Given that this study design uses simulation data, ethical approval was not required according to the Taiwan Ministry of Health and Welfare regulations.

Characteristics of the Simulated Data

The Original Survey Data

The original data were retrieved from the baseline questionnaire in the QSkin Sun and Health study [22]. A population-based cohort study of 43,794 men and women aged 40 to 69 years was randomly sampled from the population of Queensland, Australia [1], to obtain a calibration data set (two-thirds; 29,314/43,794, 66.94%) and a validation data set (one-third; 14,480/43,794, 33.06%). In the calibration data set, 24.61% (7213/29,314) of participants had a history of SC, and 75.39% (22,101/29,314) of participants did not.

The Study Simulation Data

For simplification, the 30-item difficulties calibrated in the previous study [1] (Table 1) using the Rasch partial credit model [24] were applied to yield 1000 virtual cases following a normal distribution (mean 0, SD 1; see the demonstration in Multimedia Appendix 1 with an MP4 video). The suggested cutoff point was set at 0.88 logits [1] to determine the 2 groups of cancer and noncancer in the simulation data. As such, the data with 1000 people × 30 items and 1 label (ie, 1 and 0 for melanoma status defined as cancer and noncancer groups) were applied in this study with the following 2 sections (ie, 3 models and 3 tasks).

Table 1. Overall and threshold difficulties in logit (log odds) across the 30 items.

Number	Variable	Overall difficulty	Threshold difficulty			
			Step 1	Step 2	Step 3	Step 4
1	Gender (male as 1 and female as 0)	0.16	0.00	N/A[a]	N/A	N/A
2	Skin color on areas never exposed to the sun?	−2.32	−2.46	0.78	1.68	N/A
3	Your behavior in the strong sun for 30 minutes at noon?	−0.17	−1.51	0.41	1.10	N/A
4	Your behavior outdoors in the sun without protecting your skin?	−0.49	−0.85	−0.42	1.27	N/A
5	What color are your eyes?	−0.11	−0.04	0.60	1.55	−2.11
6	What was your natural hair color at the age of 21 years?	0.48	0.70	−0.83	−1.30	1.43
7	How many freckles were on your face at the age of 21 years?	0.72	−0.37	0.01	0.36	N/A
8	How many moles did you have on your skin at the age of 21 years?	0.76	−1.45	0.53	0.92	N/A
9	How many times in your whole life have you used sunbeds?	1.27	1.35	0.30	−0.75	−0.69
10	How many separate skin cancers have you ever had excised from your skin?	0.98	0.45	−1.36	1.30	−0.39
11	How many separate sunspots or skin cancers have you ever had frozen or burnt off on your skin?	0.53	−0.05	0.49	−0.22	−0.11
12	Have I been told that I have melanoma?	0.99	0.99	N/A	N/A	N/A
13	Will you get melanoma at some point in the future?	0.26	−1.14	−0.82	1.14	0.82
14	How many times were you sunburned so badly that you were sore for at least 2 days or your skin peeled as a child?	0.58	−1.41	0.37	0.11	0.36
15	How many times were you sunburned so badly that you were sore for at least 2 days or your skin peeled in your teenage years?	0.17	−2.40	0.35	0.27	0.74
16	How many times were you sunburned so badly that you were sore for at least 2 days or your skin peeled in adulthood?	0.58	−1.83	0.59	0.10	0.46
17	How many hours did you spend outdoors and in the sun from Monday to Friday in the past year?	0.29	−0.04	0.44	−0.39	N/A
18	How many hours did you spend outdoors and in the sun from Monday to Friday at the age of 10 to 19 years?	−0.51	−0.65	0.24	0.41	N/A
19	How many hours did you spend outdoors and in the sun from Monday to Friday at the age of 20 to 29 years?	−0.15	−0.46	0.41	0.05	N/A
20	How many hours did you spend outdoors and in the sun from Monday to Friday at the age of 30 to 39 years?	0.04	−0.29	0.42	−0.13	N/A
21	How many hours did you spend outdoors and in the sun during Saturday and Sunday in the past year?	−0.14	−0.42	0.23	0.19	N/A
22	How many hours did you spend outdoors and in the sun during Saturday and Sunday at the age of 10 to 19 years?	−0.94	−0.46	0.21	0.26	N/A
23	How many hours did you spend outdoors and in the sun during Saturday and Sunday at the age of 20 to 29 years?	−0.72	−0.60	0.18	0.43	N/A
24	How many hours did you spend outdoors and in the sun during Saturday and Sunday at the age of 30 to 39 years?	−0.45	−0.56	0.19	0.37	N/A
25	Routinely apply sunscreen to my face	−0.46	0.00	N/A	N/A	N/A
26	Routinely apply sunscreen to my hands and forearms	−1.80	0.00	N/A	N/A	N/A
27	Routinely apply sunscreen to other parts of my body	−2.56	0.00	N/A	N/A	N/A
28	Routinely apply sunscreen going out in the sun: no	−0.36	0.00	N/A	N/A	N/A
29	Whether applying sunscreen outside in the sun?	−0.31	−0.90	−0.16	1.06	N/A
30	How often have you been outside in the sun in the past year?	0.45	−0.77	−0.08	0.85	N/A
31	Melanoma status (label as cancer and noncancer group)	N/A	N/A	N/A	N/A	N/A

[a]N/A: not applicable.

The 3 Models of Machine Learning Used in Microsoft Excel

The 3 Models Applied in This Study

Three models of machine learning—naïve Bayes (NB) [25], k-nearest neighbors (KNN) [26], and logistic regression (LR) [27-31]—were applied to compare the model accuracy of classifying SC in the 1000×30 rectangle data set. The 2 training (70%) and testing (30%) sets (ie, the hold-out validation) were separated to examine the model's accuracy with a proportion of 70:30, where the former was used to predict the latter.

We calculated the sensitivity, specificity, receiver operating characteristic curve (area under the curve [AUC]), and CIs along with the accuracy and precision across the 3 aforementioned models for comparison. In addition, k-fold cross-validation was performed for the 3 models using the Weka software (University of Waikato) [32]. If the Weka Explorer (graphical user interface) and the *Classify* tab are selected, we can find it by looking for the *Choose* button under the *Classify* tab. Once we navigate through the folders, the 3 classifiers are used (ie, NB classifiers→Bayes→NB; instance-based learner [IBk] classifiers→lazy→IBk; and classifiers→functions→logistic). For instance, once we select IBk for the KNN classifier, we click on the box immediately to the right of the button. This will open up a large number of options. If we then click on the button *More* in the *Options* window, we will see all the options explained. We can do this for all the classifiers to obtain additional information (eg, NB, logistic, or more; see the demonstration using an MP4 video in Multimedia Appendix 2). Meanwhile, more information about the 3 models is provided in Multimedia Appendix 3.

Calculation of Model Accuracy

After the parameters in the selected model are estimated, the accuracy of a model in the training and testing sets can be obtained through the following equations [33,34]:

The accuracy was determined by observing the higher sensitivity, specificity, precision, accuracy, and AUC in the models. The definitions are as follows:

True positive (TP) = the number of predicted cancers to the true SCs **(1)**

True negative (TN) = the number of predicted non-SCs to the true noncancers **(2)**

False positive (FP) = the number of noncancers – the number of TN **(3)**

False negative (FN) = the number of cancers – the number of TP **(4)**

Sensitivity = TP rate = TP/(TP + FN) **(5)**

Specificity = TN rate = TN/(TN + FP) **(6)**

Precision = positive predictive value = TP/(TP + FP) **(7)**

Accuracy = (TP + TN)/N **(8)**

N = TP + TN + FP + FN **(9)**

AUC = (1 – specificity) × sensitivity/2 + (sensitivity + 1) × specificity/2 **(10)**

SE for AUC = $\sqrt{(AUC \times [1 - AUC]/N)}$ **(11)**

95% CI = AUC ± 1.96 × SE for AUC **(12)**

Similarly, the confusion matrix can be made when the true conditions (ie, SC and non-SC) and the predictions (ie, positive and negative) are known in the predicted training set (or the testing data set) matched to the label (ie, 1 and 0 as cancer and noncancer groups) in the training set. Other indicators in equations (1) to (12) can be obtained accordingly.

It is worth noting that we made the model residual with the average values in the 2 groups (ie, average [range in the group of SC] + average [range in the group of non-SC]) to overcome the imbalance class data. As such, the AUC for sensitivity and specificity could be balanced in reports [35]. Details about the setting formula are provided in the Microsoft Excel module in Multimedia Appendix 1.

The 3 Tasks

Feature Variables Shown on a Forest Plot (Task 1)

The 30 variables [1] were shown on a forest plot [36-38] via the following steps: standardize each variable based on the mean (0) and SD (1) and compare the standardized mean difference on a forest plot [39].

The chi-square test was conducted to evaluate the heterogeneity between variables. Forest plots (CI plots) were drawn to display the effect estimates and their CIs for each indicator.

Comparing the Accuracies in Models (Task 2)

We calculated the sensitivity, specificity, AUC, and CIs along with the accuracy and precision across the proposed models in comparison using equations (1) to (12). Both AUCs in the training and testing sets were compared to assess the model accuracy and stability [34,35].

SC Risk and Classification (Task 3)

The Rasch Model and the First-Order Derivative in Calculus

In the Rasch model, the probability can be expressed as follows:

(13)

where θ is the person's ability, and δ is the item difficulty for person *n* and item *i*, respectively. The processes of the first-order derivative on θ are described below:

(14)

The Newton-Raphson Iteration Method

The Newton-Raphson iteration method, one of the essential iteration techniques for parameter estimation, has been frequently mentioned in the methodology literature [40-43] and popularly used in practice with the Rasch model [44,45].

A revised estimated measure, $\theta_m + 1$, is obtained from the previous measure of θ_m and the adjustment by the residual and the summed variance (defined by f'[$\theta_m - \delta_i$] across all answered items in equation 15):

(15)

The CAT SE is defined by the following equation:

(16)

The next selected item is determined by the maximum information (variance = f'[$\theta_m - \delta_i$]) of the item in all answered items shown in the following equation:

Information$_i$ = f'($\theta_m - \delta_i$) **(17)**

CAT Stop Criterion

The CAT termination is set at the CAT SE smaller than the SE of measurement (SEM) [1,46].

SEM = SD $\sqrt{(1 - Rel)}$ **(18)**

Rel is the Cronbach α of the questionnaire. Therefore, if there is a test (or questionnaire) with an SD of 1.0 logits and a Cronbach α of .78 [1], the SEM would be 0.469 ($1 \times \sqrt{[1 - .78]}$).

If CAT is terminated, the responses to unanswered items are filled in with their expected values using equation (13) when the final measure is known. The SC classification is then performed (Figure 1).

Figure 1. SC–CAT process and SC classification using machine learning. CAT: computerized adaptive testing; SC: skin cancer.
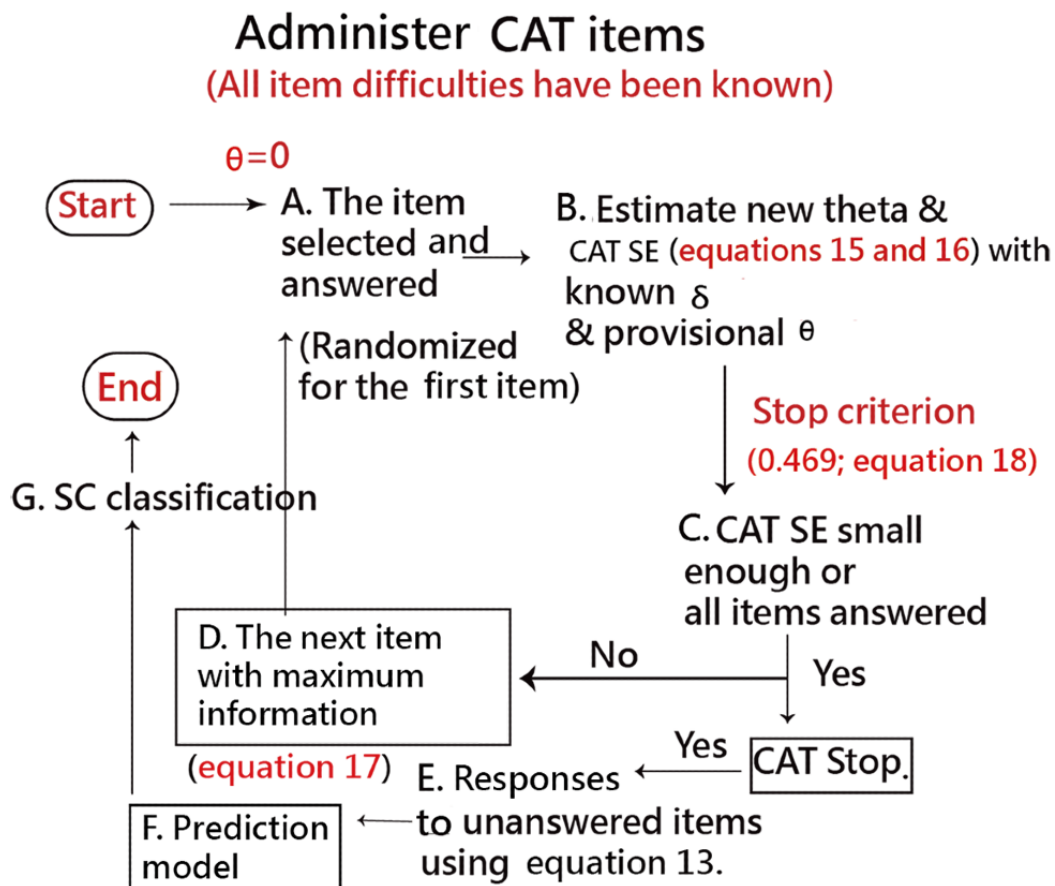
The Fit Statistics of the Mean Square Error

The Rasch fit statistics of mean square errors (MNSQs), including infit and outfit [40,41], are shown on the SC CAT to represent the extent of the deviation from the expectation of the Rasch model for the examinee's responses.

Infit MNSQ =

(19)

Outfit MNSQ =

(20)

where O_{ni} is the observed response for person n on item i, and E_{ni} is the corresponding expected value in equation (13). The variance is referred to in equations (14) and (17).

Again, another way to judge a person's responses depends on the Z score (denoted by Z) in equation (21). According to the Rasch model, these accumulated Z^2 values ought to follow a chi-square distribution with 1 degree of freedom (denoted by df) for each Z^2 value minus the degree of freedom necessary to estimate the person measure θ_n [47]. Any sum of Z^2, when divided by its df, should follow the mean square distribution in equation (22). This can conveniently be evaluated as the t statistic, which has approximately a unit normal distribution (ie, $N[0,1]$) [46], shown in equation (23).

(21)

(22)

(23)

The Skin Cancer–Computerized Adaptive Testing Algorithm

Wright [48] suggests a simpler algorithm for classroom use, classification, and performance tracking in a low-stakes environment. This algorithm is easy to implement and could be successfully used at the end of each learning module to keep track of the persons' responses in the process [46]. Figure 1 shows the core steps of skin cancer–computerized adaptive testing (SC–CAT) needed for practical adaptive testing using the Rasch model:

1. Start with a patient at an initial θ (SC score in logit) of 0.
2. Find a randomized item from the item poll via the SC–CAT.
3. Respond to the item with difficulty and the corresponding threshold δ (difficulty; label A in Figure 1).
4. Calculate the provisional θ in equation (15) based on the known item difficulties (label B).
5. Examine whether the CAT stop criterion (ie, SEM=0.469) is reached in equations (16) and (18) (label C).
6. Select the next item in equation (17) if the SC–CAT continues (label D).
7. Return to Step 3.
8. Fill in the expected values of the unanswered items via equation (13) when the SC–CAT stops based on the final estimated θ (label E).

9. Perform the prediction model (label F).
10. Obtain the classification (ie, SC or non-SC; label G).

The App Developed in This Study

An app for the detection of SC in adults was designed and developed. A 30-item self-assessment app using mobile phones was designed to predict and classify SC using machine learning and model parameters. The model parameters were embedded in the computer module.

The results of the classification (ie, SC+ and SC–) instantly appear on smartphones. A visual representation displaying the classification effect is plotted using 2 curves (ie, one from the bottom left to the top right corner denotes the success [SC+] feature, and another from the top left to the bottom right is the failure [SC–] attribute). The visual dashboard with binary (ie, SC+ and SC–) category curves is shown on Google Maps.

Statistical Tools and Data Analysis

MedCalc 9.5.0.0 for Windows (MedCalc Software) was used to calculate the sensitivity, specificity, and the corresponding AUC using LR when the observed labels (ie, 0 for SC– and 1 for SC+) and the predicted probabilities (ie, the continuous variable in equation 13) were applied.

Author-made modules in Microsoft Excel were applied to compute the model prediction indicators expressed in equations (1) to (12). The three proposed models—NB, KNN, and LR—were performed using Microsoft Excel and Weka [32] (Multimedia Appendix 1 and 2). The web-based CAT was programmed using the classic active server pages.

The study flowchart (shown in Figure 2) comprises two parts: one is from the previous study [1] and another includes 3 models. A total of 3 tasks are elaborated in this study. The abstract video is provided in Multimedia Appendix 1 as well.

Figure 2. Two major parts are in the study flowchart (in the upper and bottom panels), and three tasks are in the bottom panel. AUC: area under the curve; KNN: k-nearest neighbors; MNSQ: mean square error; SC: skin cancer; HO: hold out validation.
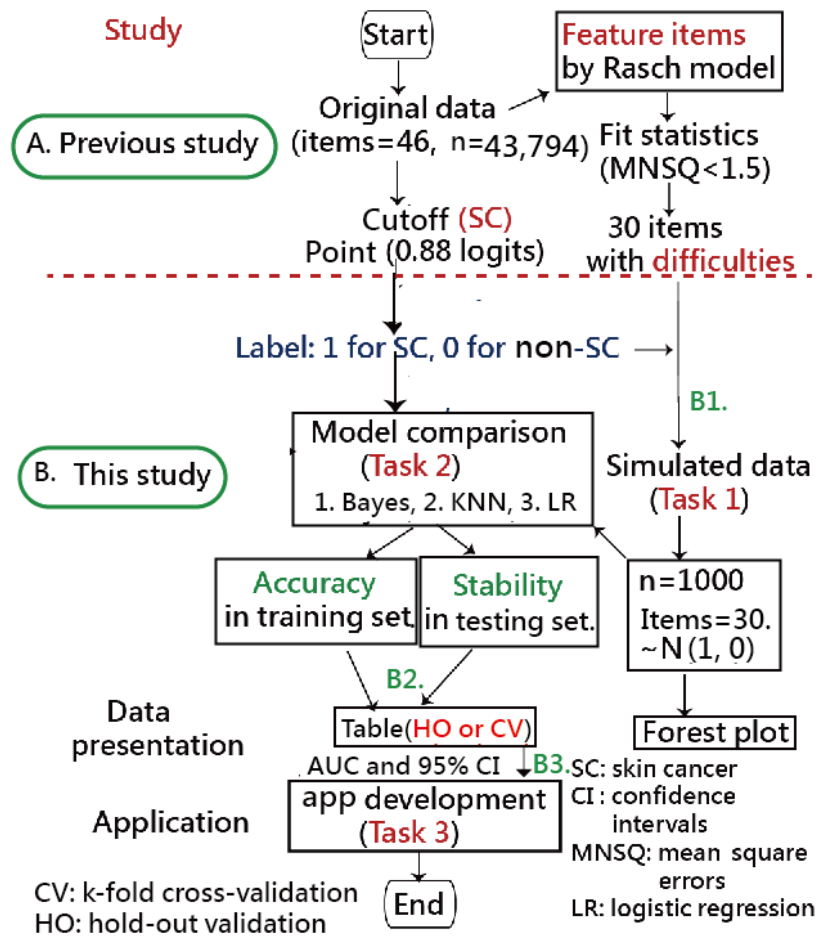
Ethics Approval and Consent to Participate

Not applicable. All data were simulated and extracted from a previous study [1].

Availability of Data and Materials

All data used in this study are available in the Multimedia Appendices.

Results

Task 1: Feature Variables Demonstrated on a Forest Plot

The 30 variables are presented in a forest plot (Figure 3). We can see that all green boxes are on the right side beyond the mean standardized mean difference (0), indicating that the variables are eligible (*P*<.05) for discriminating the melanoma status (ie, SC and non-SC groups).

Figure 3. Using the forest plot to display feature variables on smartphones [49] or clicking the QR Code. SMD: standardized mean difference.
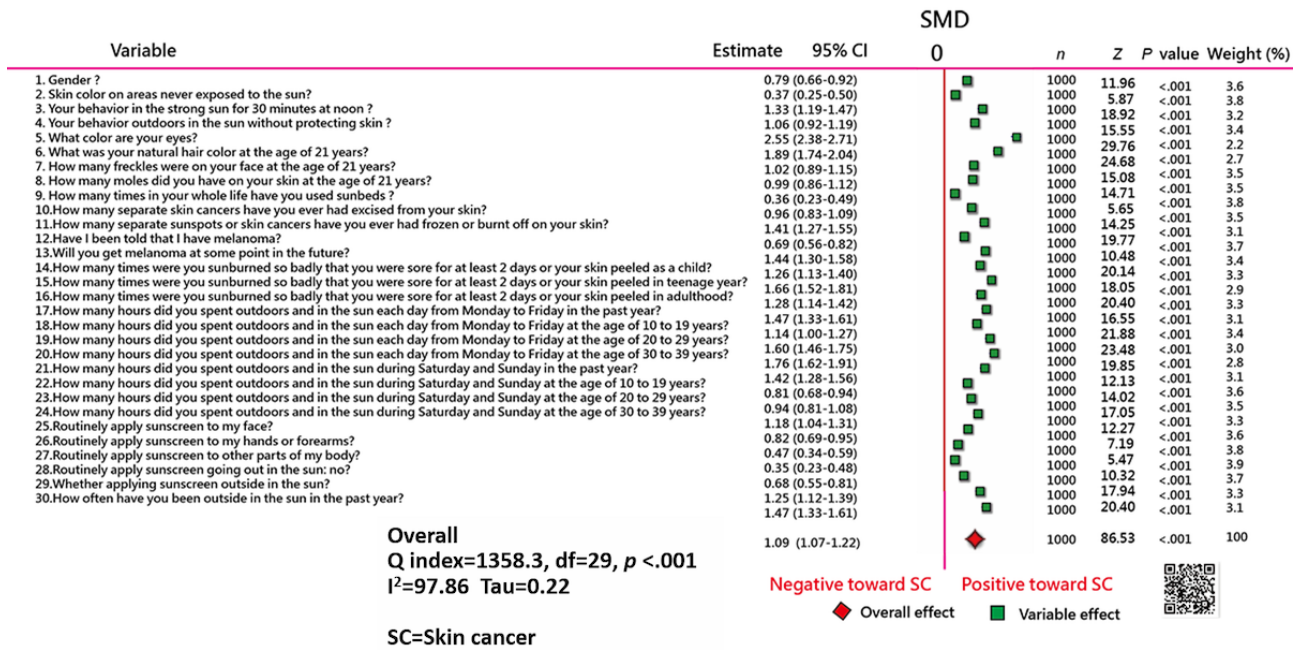
Task 2: Comparing the Accuracies Between Models

A comparison of the model accuracies is shown in Table 2. We can see that all AUCs are >0.80 in models across the training and testing sets. The 30-item KNN model yielded higher AUC values of 99% and 91% for the 700 training and 300 testing cases, respectively, far beyond the other 2 models (ie, NB and LR; Table 3). However, if k-fold cross-validation is performed, the 30-item KNN model yields lower AUC values of 85% (95% CI 83%-87%), shown in Table 4.

Table 2. Comparison of model accuracy and stability using simulation data (hold-out validation).

Study model	Training cases/testing cases, N	Accuracy ≥0.80 (training sets)					Stability ≥0.70 (testing sets)				
		Sensitivity	Specificity	Precision	Accuracy	AUC[a]	Sensitivity	Specificity	Precision	Accuracy	AUC
Naïve Bayes	700/300	0.92	0.89	0.82	0.90	0.90	0.79	0.98	0.97	0.91	0.89
KNN[b]	700/300	0.98	0.99	0.98	0.99	0.99	0.83	0.99	0.99	0.93	0.91
LR[c]	700/300	0.82	0.91	0.84	0.88	0.87	0.70	0.92	0.85	0.84	0.81

[a]AUC: area under the curve.

[b]KNN: k-nearest neighbors.

[c]LR: logistic regression.

Table 3. Comparison of model accuracy and stability using simulation data (95% CIs of the area under the curve [AUC] for hold-out validation)[a].

Study model	Accuracy ≥0.80 (training sets)			Stability ≥0.70 (testing sets)		
	Training cases, N	AUC (95% CI)	Significant difference	Testing cases, N	AUC (95% CI)	Significant difference
Naïve Bayes (1)	700	0.90 (0.88-0.92)	1, 2	300	0.89 (0.85-0.93)	__[b]
KNN[c] (2)	700	0.99 (0.98-1.00)	1, 3	300	0.91 (0.88-0.94)	3
LR[d] (3)	700	0.87 (0.85-0.89)	1, 2	300	0.81 (0.77-0.85)	2

[a]The computation of the 95% CI for the AUC is referred to in equations (10) to (12).

[b]Data not available.

[c]KNN: k-nearest neighbors.

[d]LR: logistic regression.

Table 4. Comparison of model accuracy and stability using simulation data (k-fold cross-validation).

Study model	Training cases/testing cases, N	Accuracy ≥0.80 (training sets)					Stability ≥0.70 (testing sets)	
		Sensitivity	Specificity	Precision	Accuracy	AUC[a]	AUC (95% CI)	Significant difference
Naïve Bayes (1)	700/300	0.93	0.92	0.87	92.40	0.98	0.98 (0.97-0.99)	2
KNN[b] (2)	700/300	0.87	0.90	0.84	89.20	0.85	0.85 (0.83-0.87)	1, 2
LR[c] (3)	700/300	0.90	0.90	0.90	92.40	0.98	0.98 (0.97-0.98)	2

[a]AUC: area under the curve.

[b]KNN: k-nearest neighbors.

[c]LR: logistic regression.

Task 3: Developing an App for SC Classification

A screenshot obtained from a mobile phone used to respond to the questions is shown in Figure 4, the CAT process is shown in Figure 5, and the assessment results are shown in Figure 6. In this example, we can see that the item-by-item CAT process is displayed in Figure 5, and the patient has a high probability (0.88) of developing SC, as shown in Figure 6.

Readers are invited to scan the QR code in Figure 4 and practice the web-based CAT on their own. The CAT process is shown in Figure 5. The assessment of the calibration plot is shown in Figure 6.

Figure 4. Snapshot of skin cancer assessment on smartphones from the web-based CAT model [50] or clicking the QR Code.
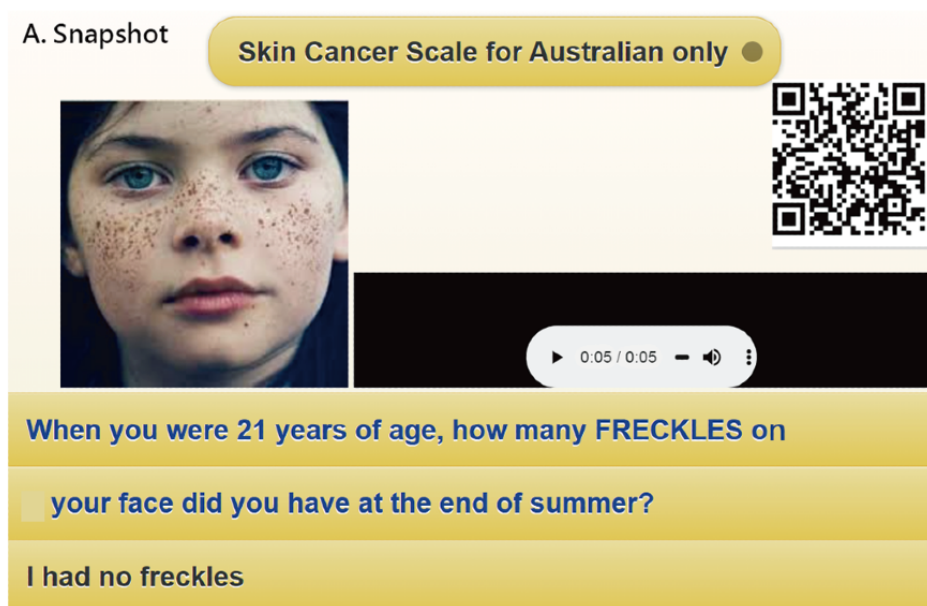
We developed the CAT-based app for classifying SC in adults. The CAT process was demonstrated item by item and is shown in the 3 panels of Figure 5. Person θ is the provisional ability (eg, the third column in the top panel of Figure 5 or the blue line in the middle panel of Figure 5) estimated by the CAT module (equation 15).

The SEs (equation 16) are along the orange line in the middle panel of Figure 5 (or the dotted lines in the top panel of Figure 5). We can see that the more items responded to by a patient, the smaller the SEs will be. The SE was generated by the formula $1/\sqrt{(\Sigma \text{information}[i])}$ (equation 17), where i refers to the CAT items responded to by a patient.

In addition, the item difficulties (shown in Table 1) are along the green line in the middle panel of Figure 5. The residual is derived from the difference (observed – expected; bottom panel of Figure 5). The Z score (i) along the brown line is computed using equation (21), which equals the squared variance (i) shown in the bottom panel of Figure 5.

CAT will stop if the residual value is <0.05. The correlation coefficient between the CAT estimated measures and the step series numbers using the last 5 estimated θ values was computed. A flatter θ trend indicates a higher probability of a person's measure converging to the final estimation.

It is worth noting that a person's MNSQs (ie, infit and outfit at the top of the middle panel in Figure 5) are generated by the formula in equations (19) and (20). If the value of the outfit is >2.0 [51], the person's response pattern is significantly aberrant beyond the model's expectation. In the example shown in the middle panel of Figure 5, we can see that the patient's response pattern with outfit MNSQ (0.52, less than the cutoff point of 2.0) and the t statistic ($-0.95 = [\ln(0.585) + 0.585 - 1] \times$,

where $v = 0.52 \times 9/[9 - 1]$ based on equations (22) and (23) meets the expectation of the Rasch model rather well.

Once the CAT terminates, the resulting example is shown in Figure 6. We can see that the SC+ with a high probability (0.88) is shown on the curve of success from the bottom left to the top right corner. The sum of both probabilities (ie, SC+ and SC–) equals 1.0. The odds can be computed by the formula $p/(1 - p) = 0.88/0.12 = 7.33$, indicating that the patient had an extremely high probability or tendency toward SC+. It is worth noting that CAT substantially reduces participant burden (ie, only 9 items were responded to in the CAT, and 70% [(30 – 9)/30] efficiency gains were from the CAT) without compromising measurement precision.

Figure 5. The process in SC-CAT on smartphones with three panels A, B, C denoted by steps, visualizations and records, respectively. CAT: computerized adaptive testing; SC: skin cancer; SEM: standard error of measurement.
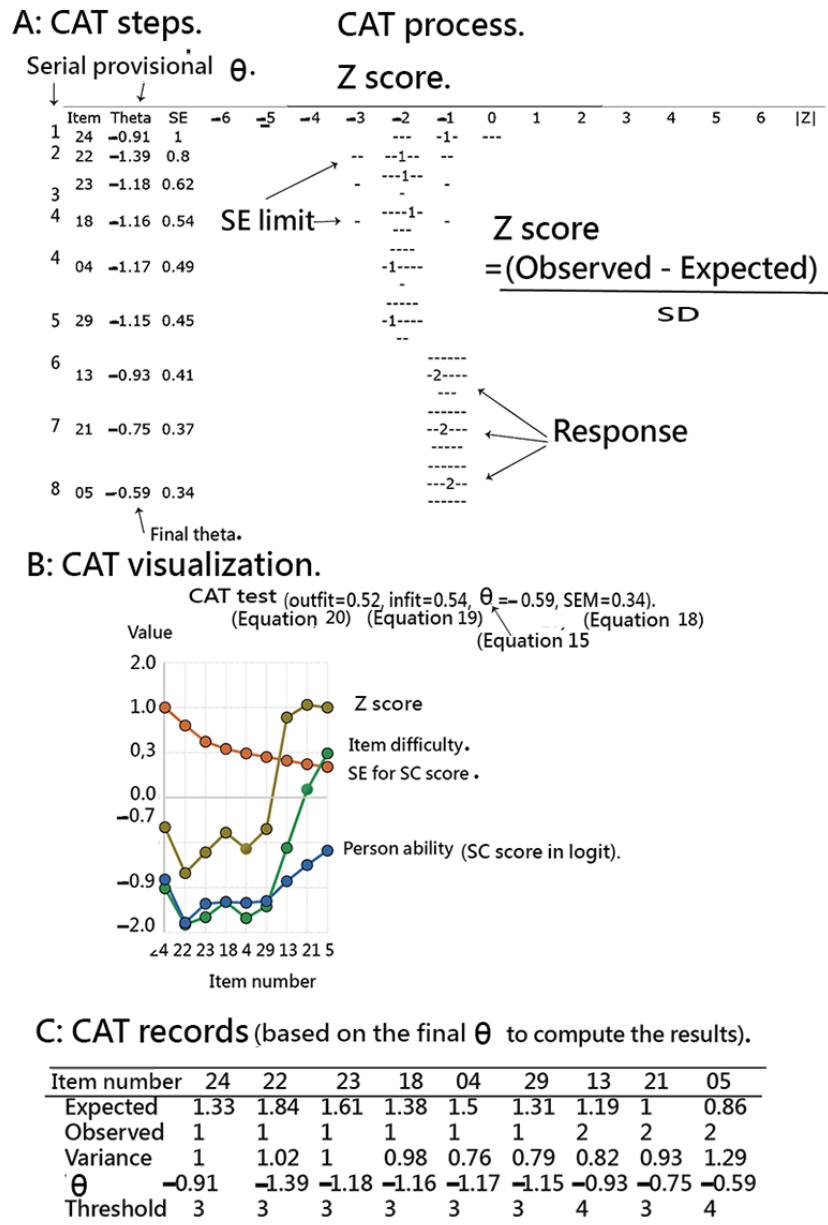
Figure 6. The result of SC+ assessment with classification and probability on smartphones. SC: skin cancer.
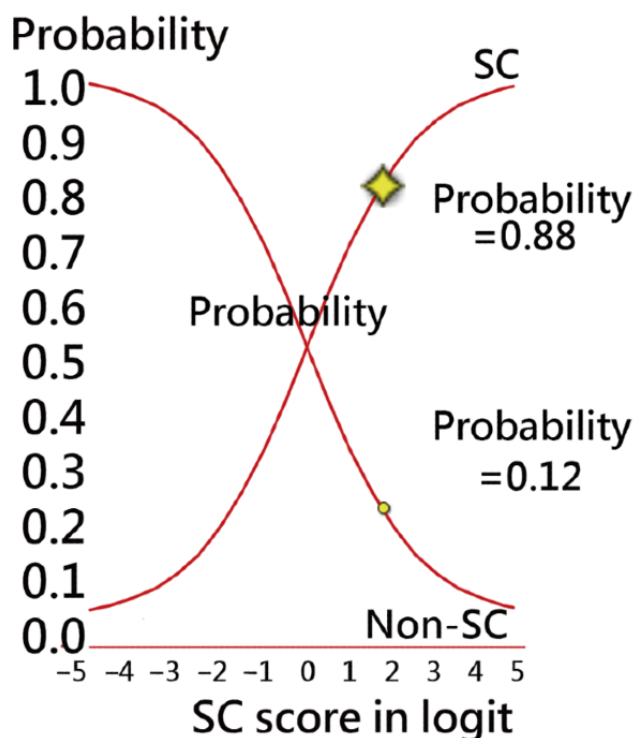
Web-Based Dashboards Shown on Google Maps

A total of 2 QR codes shown in Figures 3 and 4 (or links [49,50]) are provided for readers who can manipulate the dashboards on their own. In Figures 3 and 4, the animation-type dashboards make the data (eg, feature variables) and the app easier and clearer to understand once the QR Codes are clicked on.

Discussion

Principal Findings

We built a CAT-based model via a machine learning approach to develop an app to predict the classification of SC and help patients identify SC risk earlier to reduce participant burden and maintain acceptable measurement precision. A total of 1000 cases were simulated based on the item difficulties with a cutoff point of 0.88 logits to determine 2 groups (cancer and noncancer) using Rasch analysis addressed in a previous study [1]. A total of 3 types of machine learning (NB, KNN, and LR) were applied to compare the accuracy and stability of the models in SC classification. We observed that (1) the 30-item KNN model yielded higher AUC values of 99% and 91% for the 700 training and 300 testing cases, respectively, than its 2 counterparts using the hold-out validation but had lower AUC values of 85% (95% CI 83%-87%) in the k-fold cross-validation and (2) an app for patients that predicts SC classification was successfully developed and demonstrated in this study.

Previous Research Using Computers to Diagnose SC Instead of Classifying SC

Melanoma is considered one of the fastest-growing and most aggressive SCs; it was first described as a "fatal black tumor" by Hippocrates in 5000 BC and was later recognized to have the propensity to metastasize by William Norris in 1820 [52]. It causes most of the deaths from SC. Therefore, timely and accurate recognition of melanoma combined with appropriate treatment regimens could optimize clinical outcomes and avoid potentially fatal metastasis. Although computer-based algorithms have been proposed to develop novel predictors of prognosis and improve the efficiency and diagnostic accuracy of cancer metastasis, significant challenges for SC prediction and classification still remain [52].

For instance, a report that sniffer dogs are able to detect MM at a curable stage was first described in the United Kingdom by William et al [53]. Thereafter, studies focusing on the utility of dog olfaction for screening or diagnosing different medical conditions, such as COVID-19, malignancies, diabetes, Parkinson disease, seizures, certain hormonal and enzymatic defects [54-67], and melanoma [53], ensued. Machine learning models based on CNNs were applied to extract the region of interest of the skin lesion data set and showed that training CNN models with the region of interest–extracted data set could improve the accuracy of the prediction [55-57].

A mobile CAT was developed to help people efficiently assess their SC risk [1]. However, no such classification of SC using machine learning was provided to readers before, as we did in Figure 4 of this study. This mobile assessment could be used to quickly estimate a person's SC risk and educate patients about the need to implement skin protection and promote

self-examination of the skin [68-70]. In particular, patients with a history of SC had a higher mean score of responses than those without a history of SC [1].

Animation-Type CAT Module to Increase Health Literacy for Patients

Patients' health literacy (eg, understanding their own SC risk) is increasingly considered a critical factor affecting patient-physician communication and health outcomes [71]. Populations with below-basic or basic health literacy are less likely to obtain health issue–related information from traditional printed sources such as newspapers, magazines, books, or brochures than those with higher health literacy [72]. A brief CAT, such as the one we developed in this study, could be used to inform people quickly about their potential risk of SC and help these individuals engage in sun-protective behaviors.

This CAT module is a practical tool that can efficiently identify suitable item subsets for each individual and, therefore, maximize the efficiency and precision of the entire testing process. Through CAT, it was found that it can save up to 42% (or more) of test length and achieve a very similar degree of measurement precision as a non-CAT. This is consistent with the literature [73-76].

The tool offers diagnostics that can help practitioners assess whether responses are distorted or abnormal. For example, outfit mean square values of ≥2.0 suggest an unusual response [51]. If responses do not fit well with the model's requirement, they can be highlighted for suspected cheating, careless responding, lucky guessing, creative responding, or random responding [74]. Otherwise, one can take follow-up action (eg, medical consultation) to recheck the reasons for unexpected responses to questions [8,77,78] if the result shows a high cancer risk. Readers are invited to run the SC–CAT mobile app through the QR code, as shown in Figure 4.

Strengths and Features of This Study

There are two major forms of standardized assessments in clinical settings [79]: (1) a traditional self-administered questionnaire and (2) a rapid short-form scale [80]. Each has its own advantages and shortcomings. Traditional pencil-and-paper questionnaires require higher financial investment and have a substantial burden on respondents resulting from the following rationale: participants need to answer questions that do not provide additional information about their personal risk of certain diseases to achieve adequate precision measurement [20]. In contrast, by administering items that are most informative for the examinee, the CAT can provide precise measurement of an examinee's proficiency with the fewest possible items and then terminate at an appropriate number of items according to the required person reliability [1] (equation 18).

Second, not all questions were answered in the CAT. In contrast to those using the mean value [20] over the entire data set to fill in the missing values, we applied the expected value in the model for each unanswered question to fill in the missing data, as done in previous studies [13,24,25]. By doing so, the expected responses and model parameters can be applied to classify the SC groups. To date, we have not seen anyone using CAT

combined with machine learning to classify SC in the literature, which is a breakthrough and the second feature of this study.

Third, as with all forms of web-based technology, advances in mobile health and health communication technology are rapidly emerging [21]. The use of mobile web-based CAT is promising and worth implementing in many fields for the assessment of health issues. The CAT graphical representations shown in Figure 4 are modern and innovative in academics.

Few studies have used machine learning to perform NB, KNN, and LR on Microsoft Excel, as we did in this study. These modules are provided in Multimedia Appendix 1, which is the fourth feature of this study.

We applied the LN algorithm along with the model's parameters to design a routine on an app that is used to classify individual SCs (Figure 6), which is the fifth feature of this study. We have not seen any such SC–CAT combined with LN implemented on mobile phones before.

Different results were found when comparing the model accuracy of the AUC between the hold-out validation and the k-fold cross-validation (Tables 2, 3, and 4), which might be attributed to the small sample size (eg, 1000) used in this study. The evidence providing the k-fold cross-validation to improve the strength and confidence in the models' evaluation is the sixth feature of this study.

Limitations and Future Studies

Our study has some limitations. First, although the psychometric properties of the 30-item SC assessment have been validated for measuring SC risk [1], there is no evidence to support that the 30-item SC assessment is suitable for users outside of Australia. We recommend additional studies using their own database of SC assessment to estimate the item parameters and see whether a difference exists.

Second, although the Bayesian model performed better than the other 2 models (KNN and LR), CAT was incorporated with LR instead of the Bayesian model. The reason for this is that LR requires less computation time than the Bayes and KNN algorithms, as the latter uses pair-to-pair comparison in the algorithm. Future studies are encouraged to compare the efficiency and time consumption in computation between different models.

Third, the study was based on an article [1] that used the 30-item SC–CAT module. All the model parameters (ie, item difficulties and step-threshold difficulties) were derived from this study [1]. If any environment or condition is changed (eg, other populations in the country and different ethnicities), the result (eg, the model's parameters) will be different from that of this study. The ethnicity of the study population was also a limitation. It is worth further verifying and investigating different populations and ethnic groups under the concept we used in this study.

Fourth, the SC assessment is a 1-dimensional construct. The item difficulties used to estimate a person's measure were calibrated using Rasch Winsteps software. Traditionally, a person's ability (θ) should be estimated by the CAT method, as previous studies have done [1,10,13,16]. In this study, the

SC group should be further classified (eg, transforming the log odds to probability in LR and determining the SC group by observing the probability greater or less than 0.5). Different models applied to CAT will use disparate classification schemes. Future studies should be cautious on this matter.

Fifth, readers are encouraged to access the app by scanning the QR code in Figure 4. Professional practical apps should be further developed for Android and iOS systems in the future.

Finally, the study sample was retrieved from the baseline questionnaire in the QSkin Sun and Health study [22]. The data used in this study were simulated from item difficulties calibrated in a previous study [1]. The Rasch partial credit model [24] was used on the simulated data owing to the different number of categories across items. Further research should focus on whether the psychometric properties of the SC assessment are similar to those of this study if other IRT models are applied.

Conclusions

The contributions of this study are (1) overcoming the problem of missing responses that limit CAT development when applying the machine learning algorithm, (2) introducing 3 models available on Microsoft Excel and the k-fold cross-validation in Weka software, and (3) demonstrating an app that incorporates Rasch CAT with numerous parameters in LR.

The 30-item SC prediction model, combined with the Rasch web-based CAT, is recommended for classifying SC in adults. An app developed to help patients self-assess SC risk at an early stage is required for application in the future.

Authors' Contributions

TWC conceived and designed the study. TYY and TWC interpreted the data, and FJL monitored the process and the manuscript. TYY and TWC drafted the manuscript. All authors have read the manuscript and have approved the final manuscript.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Data deposited at OSF (Open Science Framework) research sharing platform.
[DOCX File , 15 KB - medinform_v10i3e33006_app1.docx]

Multimedia Appendix 2
K-fold cross validation performed in Weka.
[DOCX File , 14 KB - medinform_v10i3e33006_app2.docx]

Multimedia Appendix 3
Detailed information about the three models used in this study.
[DOCX File , 392 KB - medinform_v10i3e33006_app3.docx]

References

1. Djaja N, Janda M, Olsen CM, Whiteman DC, Chien T. Estimating skin cancer risk: evaluating mobile computer-adaptive testing. J Med Internet Res 2016 Jan 22;18(1):e22 [FREE Full text] [doi: 10.2196/jmir.4736] [Medline: 26800642]
2. Australian Institute of Health and Welfare. URL: http://www.aihw.gov.au/WorkArea/DownloadAsset.aspx?id=60129542353 [accessed 2022-02-06]
3. Narayanan D, Saladi R, Fox J. Ultraviolet radiation and skin cancer. Int J Dermatol 2010 Sep;49(9):978-986. [doi: 10.1111/j.1365-4632.2010.04474.x] [Medline: 20883261]
4. Global Solar UV Index: A Practical Guide. Geneva: World Health Organization; 2002.
5. Rasch G. Probabilistic Models for Some Intelligence and Attainment Tests. Denmark: Danmarks Paedagogiske Institut; 1960.
6. Lerdal A, Kottorp A, Gay CL, Grov EK, Lee KA. Rasch analysis of the Beck Depression Inventory-II in stroke survivors: a cross-sectional study. J Affect Disord 2014 Apr;158:48-52 [FREE Full text] [doi: 10.1016/j.jad.2014.01.013] [Medline: 24655764]
7. Forkmann T, Boecker M, Wirtz M, Eberle N, Westhofen M, Schauerte P, et al. Development and validation of the Rasch-based Depression Screening (DESC) using Rasch analysis and structural equation modelling. J Behav Ther Exp Psychiatry 2009 Sep;40(3):468-478. [doi: 10.1016/j.jbtep.2009.06.003] [Medline: 19589499]

8. Sauer S, Ziegler M, Schmitt M. Rasch analysis of a simplified Beck Depression Inventory. Personal Individual Differences 2013 Mar;54(4):530-535. [doi: 10.1016/j.paid.2012.10.025]

9. Chien T, Wang W, Huang S, Lai W, Chow JC. A web-based computerized adaptive testing (CAT) to assess patient perception in hospitalization. J Med Internet Res 2011 Aug 15;13(3):e61 [FREE Full text] [doi: 10.2196/jmir.1785] [Medline: 21844001]

10. Ma S, Chien T, Wang H, Li Y, Yui M. Applying computerized adaptive testing to the Negative Acts Questionnaire-Revised: Rasch analysis of workplace bullying. J Med Internet Res 2014 Feb 17;16(2):e50 [FREE Full text] [doi: 10.2196/jmir.2819] [Medline: 24534113]

11. Djaja N, Youl P, Aitken J, Janda M. Evaluation of a skin self examination attitude scale using an item response theory model approach. Health Qual Life Outcomes 2014 Dec 24;12:189 [FREE Full text] [doi: 10.1186/s12955-014-0189-x] [Medline: 25539671]

12. Bjorner JB, Kreiner S, Ware JE, Damsgaard MT, Bech P. Differential item functioning in the Danish translation of the SF-36. J Clin Epidemiol 1998 Nov;51(11):1189-1202. [doi: 10.1016/s0895-4356(98)00111-5] [Medline: 9817137]

13. Ma S, Chou W, Chien T, Chow JC, Yeh Y, Chou P, et al. An app for detecting bullying of nurses using convolutional neural networks and web-based computerized adaptive testing: development and usability study. JMIR Mhealth Uhealth 2020 May 20;8(5):e16747 [FREE Full text] [doi: 10.2196/16747] [Medline: 32432557]

14. Lord FM. Practical applications of item characteristic curve theory. J Educ Measurement 1977 Jun;14(2):117-138. [doi: 10.1111/j.1745-3984.1977.tb00032.x]

15. Lord F. Applications of Item Response Theory To Practical Testing Problems. Milton Park, Abingdon-on-Thames, Oxfordshire United Kingdom: Taylor & Francis; 1980.

16. Ma S, Wang H, Chien T. A new technique to measure online bullying: online computerized adaptive testing. Ann Gen Psychiatry 2017;16:26 [FREE Full text] [doi: 10.1186/s12991-017-0149-z] [Medline: 28680455]

17. Lee Y, Chou W, Chien T, Chou P, Yeh Y, Lee H. An app developed for detecting nurse burnouts using the convolutional neural networks in Microsoft excel: population-based questionnaire study. JMIR Med Inform 2020 May 07;8(5):e16528 [FREE Full text] [doi: 10.2196/16528] [Medline: 32379050]

18. Chien T, Lin W. Simulation study of activities of daily living functions using online computerized adaptive testing. BMC Med Inform Decis Mak 2016 Oct 10;16(1):130 [FREE Full text] [doi: 10.1186/s12911-016-0370-8] [Medline: 27724939]

19. Ravi D, Wong C, Deligianni F, Berthelot M, Andreu-Perez J, Lo B, et al. Deep learning for health informatics. IEEE J Biomed Health Inform 2017 Jan;21(1):4-21. [doi: 10.1109/JBHI.2016.2636665] [Medline: 28055930]

20. Wang H, Cui Z, Chen Y, Avidan M, Abdallah AB, Kronzer A. Predicting hospital readmission via cost-sensitive deep learning. IEEE/ACM Trans Comput Biol Bioinf 2018 Nov 1;15(6):1968-1978. [doi: 10.1109/tcbb.2018.2827029]

21. Mitchell SJ, Godoy L, Shabazz K, Horn IB. Internet and mobile technology use among urban African American parents: survey study of a clinical population. J Med Internet Res 2014 Jan 13;16(1):e9 [FREE Full text] [doi: 10.2196/jmir.2673] [Medline: 24418967]

22. Olsen CM, Green AC, Neale RE, Webb PM, Cicero RA, Jackman LM, QSkin Study. Cohort profile: the QSkin sun and health study. Int J Epidemiol 2012 Aug;41(4):929-92i. [doi: 10.1093/ije/dys107] [Medline: 22933644]

23. Lai P, Chien T. The determination of inflection curve on a given ogive curve using the second order derivative in calculus. J Bibliographical Analyses Stat 2021;18(3):31-33 [FREE Full text]

24. Masters GN. A rasch model for partial credit scoring. Psychometrika 1982 Jun;47(2):149-174. [doi: 10.1007/BF02296272]

25. Tang X, Shu Y, Liu W, Li J, Liu M, Yu H. An optimized weighted naïve Bayes method for flood risk assessment. Risk Anal 2021 Dec;41(12):2301-2321. [doi: 10.1111/risa.13743] [Medline: 33928661]

26. Viana Dos Santos Santana Í, Cm da Silveira A, Sobrinho A, Chaves E Silva L, Dias da Silva L, Santos DF, et al. Classification models for COVID-19 test prioritization in Brazil: machine learning approach. J Med Internet Res 2021 Apr 08;23(4):e27293 [FREE Full text] [doi: 10.2196/27293] [Medline: 33750734]

27. Golpour P, Ghayour-Mobarhan M, Saki A, Esmaily H, Taghipour A, Tajfard M, et al. Comparison of support vector machine, naïve Bayes and logistic regression for assessing the necessity for coronary angiography. Int J Environ Res Public Health 2020 Sep 04;17(18):6449 [FREE Full text] [doi: 10.3390/ijerph17186449] [Medline: 32899733]

28. Gholizadeh P, Esmaeili B. Developing a multi-variate logistic regression model to analyze accident scenarios: case of electrical contractors. Int J Environ Res Public Health 2020 Jul 06;17(13):4852 [FREE Full text] [doi: 10.3390/ijerph17134852] [Medline: 32640549]

29. Nhu V, Shirzadi A, Shahabi H, Singh SK, Al-Ansari N, Clague JJ, et al. Shallow landslide susceptibility mapping: a comparison between logistic model tree, logistic regression, naïve bayes tree, artificial neural network, and support vector machine algorithms. Int J Environ Res Public Health 2020 Apr 16;17(8):2749 [FREE Full text] [doi: 10.3390/ijerph17082749] [Medline: 32316191]

30. Choi Y, Boo Y. Comparing logistic regression models with alternative machine learning methods to predict the risk of drug intoxication mortality. Int J Environ Res Public Health 2020 Jan 31;17(3):897 [FREE Full text] [doi: 10.3390/ijerph17030897] [Medline: 32023993]

31. Wu L, Deng F, Xie Z, Hu S, Shen S, Shi J, et al. Spatial analysis of severe fever with thrombocytopenia syndrome virus in china using a geographically weighted logistic regression model. Int J Environ Res Public Health 2016 Nov 11;13(11):1125 [FREE Full text] [doi: 10.3390/ijerph13111125] [Medline: 27845737]

XSL•FO
RenderX

32. Hall M, Frank E, Holmes G, Pfahringer B, Reutemann P, Witten I. The WEKA data mining software. SIGKDD Explor Newsl 2009 Nov 16;11(1):10-18 [FREE Full text] [doi: 10.1145/1656274.1656278]

33. Rere LM, Fanany MI, Arymurthy AM. Metaheuristic algorithms for convolution neural network. Comput Intell Neurosci 2016;2016:1537325 [FREE Full text] [doi: 10.1155/2016/1537325] [Medline: 27375738]

34. Chou P, Chien T, Yang T, Yeh Y, Chou W, Yeh C. Predicting active NBA players most likely to be inducted into the basketball hall of famers using artificial neural networks in Microsoft excel: development and usability study. Int J Environ Res Public Health 2021 Apr 16;18(8):4256 [FREE Full text] [doi: 10.3390/ijerph18084256] [Medline: 33923846]

35. Tey S, Liu C, Chien T, Hsu C, Chan K, Chen C, et al. Predicting the 14-day hospital readmission of patients with pneumonia using artificial neural networks (ANN). Int J Environ Res Public Health 2021 May 12;18(10):5110 [FREE Full text] [doi: 10.3390/ijerph18105110] [Medline: 34065894]

36. Hamling J, Lee P, Weitkunat R, Ambühl M. Facilitating meta-analyses by deriving relative effect and precision estimates for alternative comparisons from a set of estimates presented by exposure level or disease category. Stat Med 2008 Mar 30;27(7):954-970. [doi: 10.1002/sim.3013] [Medline: 17676579]

37. Chen C, Wang L, Kuo H, Fang Y, Lee H. Significant effects of late evening snack on liver functions in patients with liver cirrhosis: a meta-analysis of randomized controlled trials. J Gastroenterol Hepatol 2019 Jul;34(7):1143-1152. [doi: 10.1111/jgh.14665] [Medline: 30883904]

38. Lalkhen AG, McCluskey A. Statistics V: introduction to clinical trials and systematic reviews. Continuing Educ Anaesthesia Critical Care Pain 2008 Aug;8(4):143-146. [doi: 10.1093/bjaceaccp/mkn023]

39. Yan Y, Chien T. The use of forest plot to identify article similarity and differences in characteristics between journals using medical subject headings terms: a protocol for bibliometric study. Medicine (Baltimore) 2021 Feb 12;100(6):e24610 [FREE Full text] [doi: 10.1097/MD.0000000000024610] [Medline: 33578568]

40. Wright BD, Douglas GA. Conditional versus unconditional procedures for sample-free item analysis. Educ Psychol Measurement 2016 Jul 02;37(3):573-586. [doi: 10.1177/001316447703700301]

41. Wright B, Douglas G. Estimating Rasch (person, ability, theta) measures with known dichotomous item difficulties: anchored maximum likelihood estimation (AMLE). Rasch Measurement Transactions. URL: https://www.rasch.org/rmt/rmt102t.htm [accessed 2022-02-06]

42. Ludlow L, Haley K. Newton: pinball wizard? Popular Measure 1999;2(1):5 [FREE Full text]

43. Wright BD, Stone MH. Measurement Essentials 2nd Edition. Wilmington, Delaware: Wide Range, Inc; 1999.

44. Chien T, Shao Y. Rasch analysis for continuous variables. Rasch Measurement Transact 2016;30(1):1574-1576.

45. Chien T, Shao Y, Kuo S. Development of a Microsoft Excel tool for one-parameter Rasch model of continuous items: an application to a safety attitude survey. BMC Med Res Methodol 2017 Jan 10;17(1):4 [FREE Full text] [doi: 10.1186/s12874-016-0276-2] [Medline: 28068901]

46. Linacre J. Computer-adaptive testing: a methodology whose time has come. MESA Memorandum. URL: https://www.rasch.org/memo69.htm [accessed 2022-02-06]

47. Wright B, Stone M. Best Test Design Rasch Measurement. Chicago, IL: Mesa Press; 1979.

48. Wright B. Practical adaptive testing. Rasch Measurement Transact 1988;2(2):21.

49. Chien T. iHELP system. URL: http://www.healthup.org.tw/gps/skincancer2021.htm [accessed 2022-02-06]

50. Web-based computerized adaptive testing model for skin cancer assessment on smartphones. iHELP. URL: http://www.healthup.org.tw/irs/irsin_e.asp?type1=15 [accessed 2022-02-06]

51. Linacre J. Optimizing rating scale category effectiveness. J Appl Meas 2002;3(1):85-106. [Medline: 11997586]

52. Alix-Panabieres C, Magliocco A, Cortes-Hernandez LE, Eslami- S, Franklin D, Messina JL. Detection of cancer metastasis: past, present and future. Clin Exp Metastasis 2021 May 07 (forthcoming). [doi: 10.1007/s10585-021-10088-w] [Medline: 33961169]

53. Williams H, Pembroke A. Sniffer dogs in the melanoma clinic? Lancet 1989 Apr 01;1(8640):734. [doi: 10.1016/s0140-6736(89)92257-5] [Medline: 2564551]

54. Eskandari E, Ahmadi Marzaleh M, Roudgari H, Hamidi Farahani R, Nezami-Asl A, Laripour R, et al. Sniffer dogs as a screening/diagnostic tool for COVID-19: a proof of concept study. BMC Infect Dis 2021 Mar 05;21(1):243 [FREE Full text] [doi: 10.1186/s12879-021-05939-6] [Medline: 33673823]

55. Boedeker E, Friedel G, Walles T. Sniffer dogs as part of a bimodal bionic research approach to develop a lung cancer screening. Interact Cardiovasc Thorac Surg 2012 May;14(5):511-515 [FREE Full text] [doi: 10.1093/icvts/ivr070] [Medline: 22345057]

56. Zanddizari H, Nguyen N, Zeinali B, Chang JM. A new preprocessing approach to improve the performance of CNN-based skin lesion classification. Med Biol Eng Comput 2021 May;59(5):1123-1131. [doi: 10.1007/s11517-021-02355-5] [Medline: 33904008]

57. Ningrum DN, Yuan S, Kung W, Wu C, Tzeng I, Huang C, et al. Deep learning classifier with patient's metadata of dermoscopic images in malignant melanoma detection. J Multidiscip Healthc 2021;14:877-885 [FREE Full text] [doi: 10.2147/JMDH.S306284] [Medline: 33907414]

58. Alheejawi S, Berendt R, Jha N, Maity SP, Mandal M. Automated proliferation index calculation for skin melanoma biopsy images using machine learning. Comput Med Imaging Graph 2021 Apr;89:101893. [doi: 10.1016/j.compmedimag.2021.101893] [Medline: 33752078]

59. Welsh JS. Olfactory detection of human bladder cancer by dogs: another cancer detected by "pet scan". BMJ 2004 Nov 27;329(7477):1286-1287 [FREE Full text] [doi: 10.1136/bmj.329.7477.1286-b] [Medline: 15564264]

60. Urbanová L, Vyhnánková V, Krisová S, Pacík D, Nečas A. Intensive training technique utilizing the dog's olfactory abilities to diagnose prostate cancer in men. Acta Vet Brno 2015 Mar 19;84(1):77-82. [doi: 10.2754/avb201585010077]

61. Lippi G, Cervellin G. Canine olfactory detection of cancer versus laboratory testing: myth or opportunity? Clin Chem Lab Med 2012 Mar;50(3):435-439. [doi: 10.1515/CCLM.2011.672] [Medline: 21790506]

62. Elliker K, Williams H. Detection of skin cancer odours using dogs: a step forward in melanoma detection training and research methodologies. Br J Dermatol 2016 Nov;175(5):851-852. [doi: 10.1111/bjd.15030] [Medline: 27790682]

63. Willis CM, Church SM, Guest CM, Cook WA, McCarthy N, Bransbury AJ, et al. Olfactory detection of human bladder cancer by dogs: proof of principle study. BMJ 2004 Sep 25;329(7468):712 [FREE Full text] [doi: 10.1136/bmj.329.7468.712] [Medline: 15388612]

64. Kane E. Cancer-sniffing dogs: how canine scent detection could transform human medicine. dvm360. URL: https://www.dvm360.com/view/cancer-sniffing-dogs-how-canine-scent-detection-could-transform-human-medicine [accessed 2022-02-06]

65. McCulloch M, Jezierski T, Broffman M, Hubbard A, Turner K, Janecki T. Diagnostic accuracy of canine scent detection in early- and late-stage lung and breast cancers. Integr Cancer Ther 2006 Mar;5(1):30-39 [FREE Full text] [doi: 10.1177/1534735405285096] [Medline: 16484712]

66. Ehmann R, Boedeker E, Friedrich U, Sagert J, Dippon J, Friedel G, et al. Canine scent detection in the diagnosis of lung cancer: revisiting a puzzling phenomenon. Eur Respir J 2012 Mar;39(3):669-676 [FREE Full text] [doi: 10.1183/09031936.00051711] [Medline: 21852337]

67. Los EA, Ramsey KL, Guttmann-Bauman I, Ahmann AJ. Reliability of trained dogs to alert to hypoglycemia in patients with type 1 diabetes. J Diabetes Sci Technol 2017 May;11(3):506-512 [FREE Full text] [doi: 10.1177/1932296816666537] [Medline: 27573791]

68. Robinson JK, Gaber R, Hultgren B, Eilers S, Blatt H, Stapleton J, et al. Skin self-examination education for early detection of melanoma: a randomized controlled trial of Internet, workbook, and in-person interventions. J Med Internet Res 2014 Jan 13;16(1):e7 [FREE Full text] [doi: 10.2196/jmir.2883] [Medline: 24418949]

69. Brady MS, Oliveria SA, Christos PJ, Berwick M, Coit DG, Katz J, et al. Patterns of detection in patients with cutaneous melanoma. Cancer 2000 Jul 15;89(2):342-347. [doi: 10.1002/1097-0142(20000715)89:2<342::aid-cncr19>3.0.co;2-p]

70. Berwick M, Begg CB, Fine JA, Roush GC, Barnhill RL. Screening for cutaneous melanoma by skin self-examination. J Natl Cancer Inst 1996 Jan 03;88(1):17-23. [doi: 10.1093/jnci/88.1.17] [Medline: 8847720]

71. Williams MV, Davis T, Parker RM, Weiss BD. The role of health literacy in patient-physician communication. Fam Med 2002 May;34(5):383-389. [Medline: 12038721]

72. Cutilli C, Bennett I. Understanding the health literacy of America: results of the National Assessment of Adult Literacy. Orthop Nurs 2009;28(1):27-32; quiz 33 [FREE Full text] [doi: 10.1097/01.NOR.0000345852.22122.d6] [Medline: 19190475]

73. Pedersen PM, Jørgensen HS, Nakayama H, Raaschou HO, Olsen TS. Comprehensive assessment of activities of daily living in stroke. The Copenhagen stroke study. Archives Physical Med Rehab 1997 Feb;78(2):161-165. [doi: 10.1016/s0003-9993(97)90258-6]

74. Wainer H, Dorans N, Flaugher R, Green B, Mislevy R. Computerized Adaptive Testing A Primer. Milton Park, Abingdon-on-Thames, Oxfordshire United Kingdom: Taylor & Francis; 1990.

75. Weiss DJ, McBride JR. Bias and information of bayesian adaptive testing. Applied Psychol Measure 2016 Jul 27;8(3):273-285. [doi: 10.1177/014662168400800303]

76. Chien T, Wu H, Wang W, Castillo R, Chou W. Reduction in patient burdens with graphical computerized adaptive testing on the ADL scale: tool development and simulation. Health Qual Life Outcomes 2009 May 05;7:39 [FREE Full text] [doi: 10.1186/1477-7525-7-39] [Medline: 19416521]

77. Eack SM, Singer JB, Greeno CG. Screening for anxiety and depression in community mental health: the beck anxiety and depression inventories. Community Ment Health J 2008 Dec;44(6):465-474. [doi: 10.1007/s10597-008-9150-y] [Medline: 18516678]

78. Shear MK, Greeno C, Kang J, Ludewig D, Frank E, Swartz HA, et al. Diagnosis of nonpsychotic patients in community clinics. Am J Psychiatry 2000 Apr;157(4):581-587. [doi: 10.1176/appi.ajp.157.4.581] [Medline: 10739417]

79. Ramirez Basco M, Bostic JQ, Davies D, Rush AJ, Witte B, Hendrickse W, et al. Methods to improve diagnostic accuracy in a community mental health setting. Am J Psychiatry 2000 Oct;157(10):1599-1605. [doi: 10.1176/appi.ajp.157.10.1599] [Medline: 11007713]

80. De Beurs DP, de Vries AL, de Groot MH, de Keijser J, Kerkhof AJ. Applying computer adaptive testing to optimize online assessment of suicidal behavior: a simulation study. J Med Internet Res 2014 Sep 11;16(9):e207 [FREE Full text] [doi: 10.2196/jmir.3511] [Medline: 25213259]

Abbreviations

AUC: area under the curve
CAT: computerized adaptive testing
CNN: convolutional neural network
FN: false negative
FP: false positive
IBk: instance-based learner
IRT: item response theory
KNN: k-nearest neighbors
LR: logistic regression
MM: malignant melanoma
MNSQ: mean square error
NB: naïve Bayes
NMSC: nonmelanoma skin cancer
SC: skin cancer
SC–CAT: skin cancer–computerized adaptive testing
SEM: standard error of measurement
TN: true negative
TP: true positive

XSL•FO
RenderX

<u>Original Paper</u>

Selective Prediction With Long Short-term Memory Using Unit-Wise Batch Standardization for Time Series Health Data Sets: Algorithm Development and Validation

Borum Nam[1], BS; Joo Young Kim[2], BS; In Young Kim[2*], MD; Baek Hwan Cho[3*], PhD

[1]Department of Electronic Engineering, Hanyang University, Seoul, Republic of Korea

[2]Department of Biomedical Engineering, Hanyang University, Seoul, Republic of Korea

[3]Medical AI Research Center, Samsung Medical Center, Seoul, Republic of Korea

[*]these authors contributed equally

Corresponding Author:
Baek Hwan Cho, PhD
Medical AI Research Center
Samsung Medical Center
81, Irwon-ro, Gangnam-gu
Seoul, 06351
Republic of Korea
Phone: 82 234100885
Email: baekhwan.cho@samsung.com

Abstract

Background: In any health care system, both the classification of data and the confidence level of such classifications are important. Therefore, a selective prediction model is required to classify time series health data according to confidence levels of prediction.

Objective: This study aims to develop a method using long short-term memory (LSTM) models with a reject option for time series health data classification.

Methods: An existing selective prediction method was adopted to implement an option for rejecting a classification output in LSTM models. However, a conventional selection function approach to LSTM does not achieve acceptable performance during learning stages. To tackle this problem, we proposed a unit-wise batch standardization that attempts to normalize each hidden unit in LSTM to apply the structural characteristics of LSTM models that concern the selection function.

Results: The ability of our method to approximate the target confidence level was compared by coverage violations for 2 time series of health data sets consisting of human activity and arrhythmia. For both data sets, our approach yielded lower average coverage violations (0.98% and 1.79% for each data set) than those of the conventional approach. In addition, the classification performance when using the reject option was compared with that of other normalization methods. Our method demonstrated superior performance for selective risk (12.63% and 17.82% for each data set), false-positive rates (2.09% and 5.8% for each data set), and false-negative rates (10.58% and 17.24% for each data set).

Conclusions: Our normalization approach can help make selective predictions for time series health data. We expect this technique to enhance the confidence of users in classification systems and improve collaborative efforts between humans and artificial intelligence in the medical field through the use of classification that considers confidence.

KEYWORDS

XSL•FO
RenderX

Introduction

Background

High-performance networks have been used to enhance the quality and convenience of human life since the development of deep learning techniques. Deep learning networks are used in education, aviation, process management, entertainment, agriculture, and robotics. Artificial intelligence (AI) has made significant contributions to a variety of medical applications [1-3]. However, in a clinical setting, the output from AI as an accurate prediction is often insufficient and requires its interpretation for further decisions [4]. As medical AI systems can support efficient and accurate decisions, it is important not only to increase the accuracy of classification in deep learning networks but also to reduce errors, particularly those that can be fatal [5]. In addition, health care data tend to be complex, and neural networks have proven problematic in accurately recognizing patterns in this complexity [6]. The uncertainty of prediction measures the reliability of a prediction and must be considered in fields that require prudent decisions, such as medicine or autonomous driving [7]. Accordingly, in fields where minor errors can cause significant problems, applying a prediction model that can reject predictions when the confidence level is not high enough is helpful. To develop such a deep neural network, a selective prediction [8] method can be applied to use the confidence level in both training and test sessions.

Various biosignal sensors have been developed for human health care applications, and many algorithms have been developed to analyze the data produced by these sensors. Deep learning technologies have performed well when applied to data obtained from health care or medical sensors [9]. Classification models based on a deep neural network or convolutional neural network (CNN) have been used to classify health and medical data. In addition, biosignals and time series data from humans are used in diverse health care systems [10]. In various studies, recurrent neural network (RNN) models have been used to classify health and medical data, especially time series data. Among such models, RNNs have contributed significantly to the classification of time series data. Many studies have used RNN models to classify electronic health records obtained from clinical measurements [11], predict diseases using patient diagnostic histories [12-14], conduct health status analyses using biosignals [15-18], and classify health information from mobile and wearable sensors [19-22]. Previous studies have applied prediction confidence to classify image data, and prediction confidence can be considered for classifying time series health data using RNN models. However, little research has focused on how to use prediction confidence for time series health data.

Considering the specificity of time series health data, a model that can produce results according to the predicted confidence level and uses prediction confidence has the advantage of reducing fatal errors.

The selective prediction model can learn from certain samples that are sufficiently confident in their predictions. This means that such a model can ignore predictions when they are uncertain in training. In addition, the selective prediction model provides a confidence level for each test sample in the inference stage, which can be used as a reference score in a medical situation. In early studies on selective prediction, neural network models with a reject option were used to obtain a specific confidence score from a trained model and as a model threshold to validate performance [23-25]. However, these methods calculate the prediction probability to select samples for training based on a threshold called the prediction confidence score.

Recently, research using the selective prediction model mainly consists of 2 parts. The first is to extract an appropriate prediction confidence score and the second is to make good use of the extracted prediction confidence score for the deep learning model. For extracting the prediction confidence score, methods have been designed in many studies. For example, the softmax response and Monte Carlo (MC) dropout methods use a confidence score from neural networks [26]. The softmax response method extracts a confidence score using maximum softmax values from neural networks, as described in the above methods, whereas an MC dropout estimates a confidence score using statistical approaches. However, MC dropout requires a high computational cost to optimize the problem quickly. Although Bayesian methods [27-29] can produce prediction confidence scores of RNNs [30], they are applicable only for natural language processing, which uses *many-to-many* RNNs with multiple sequence inputs and outputs. However, the predictive models in health care are usually *many-to-one* types that predict class using a health information time series as input, and it is helpful for medical staff to train a *many-to-one* predictive model for time series data that has a selective prediction ability. For a model using the prediction confidence score, a selective prediction model that learned both prediction and selection was developed [31]. On the basis of this method, SelectiveNet [32] has demonstrated potential possibilities for various applications, with the advantage of learning the selection and prediction simultaneously. However, the structure of the selective prediction model using long short-term memory (LSTM) has not been validated in previous studies. Thus, a well-designed selective prediction model for time series data is required.

Objective

In this study, a selective prediction model using LSTM [33] was implemented to classify time series health data. In particular, we considered a method that incorporates a reject option to control and measure prediction confidence for *many-to-one* classification tasks. As the selection function uses the output of the prediction model as an input, a suitable selection function structure must be devised. Therefore, methods to normalize the selection function were compared to achieve a structure suitable for classifying time series data with LSTM. To validate the LSTM selective prediction performance, we used coverage violations and selective risks for each data set. As high false-positive and false-negative rates can be critical factors in diagnoses, we also present the false-positive and false-negative rates of the LSTM selective prediction model. In summary, the goal of this study is to develop a selective prediction model for health data time series. The contributions of this study are (1) applying the latest selective prediction method with superior performance to classify time series health data using LSTM and (2) presenting the structure of the selection

function in the selective prediction model (especially the normalization method) for time series selective prediction.

Methods

Selective Prediction

We examined the possibility of RNN models with a reject option using SelectiveNet [32], which has superior performance compared with existing selective prediction models. The overall structure of the model was based on the SelectiveNet [32] model with an LSTM; it is divided into selective and auxiliary predictions, as shown in Figure 1. The selective prediction is divided again into two steps: prediction and selection. Prediction involves the results of the LSTM model and the selection part extracts the predicted confidence level of the LSTM model. In this study, we propose unit-wise batch standardization (UBS) as part of the selection function. Selective prediction is performed using both the prediction and selection function results. An auxiliary prediction step using the LSTM prediction result to derive the final result with the selective prediction result was added to enhance prediction performance. As selective prediction is a prediction model using a deep learning model structure, it is optimized by a loss function. The entire model is trained by optimizing the selective prediction and auxiliary prediction steps simultaneously. Further details are provided in the *Optimization* section. LSTM was used for the RNN model for time series data classification.

A selective model was used to implement classification models with the reject option [34]. The selective model (f, g) consists of pairing a prediction function f and a selection function $g: X \rightarrow Y$

$\{Y | 0 \leq Y \leq 1\}$ (X is a set of inputs and Y is a set of outputs). When the data set is given as ☒ for supervised learning of the classification model, the empirical risk of prediction function f becomes ☒. When τ is a threshold, g acts as a qualifier of f and can be expressed as follows:

☒

Selective models can be controlled by coverage and risk values. When E_p is the expected probability, and ℓ is the loss function, we can define the coverage and risk as follows:

☒

where $g(x)$ is the prediction confidence score, $\varphi(g)$ is a coverage value that is the expected value of the prediction confidence scores for training samples, which is correlated with the number of selected samples during training. $R(f, g)$ is a selective risk that represents the error rate for predicting the selected samples using selective prediction. The corresponding selective risk for a data set ☒ is called the empirical selective risk and is defined as follows:

☒

The empirical coverage corresponding to the data set S_m is as follows:

☒

Figure 1. Long short-term memory model structure with a reject option. LSTM: long short-term memory.
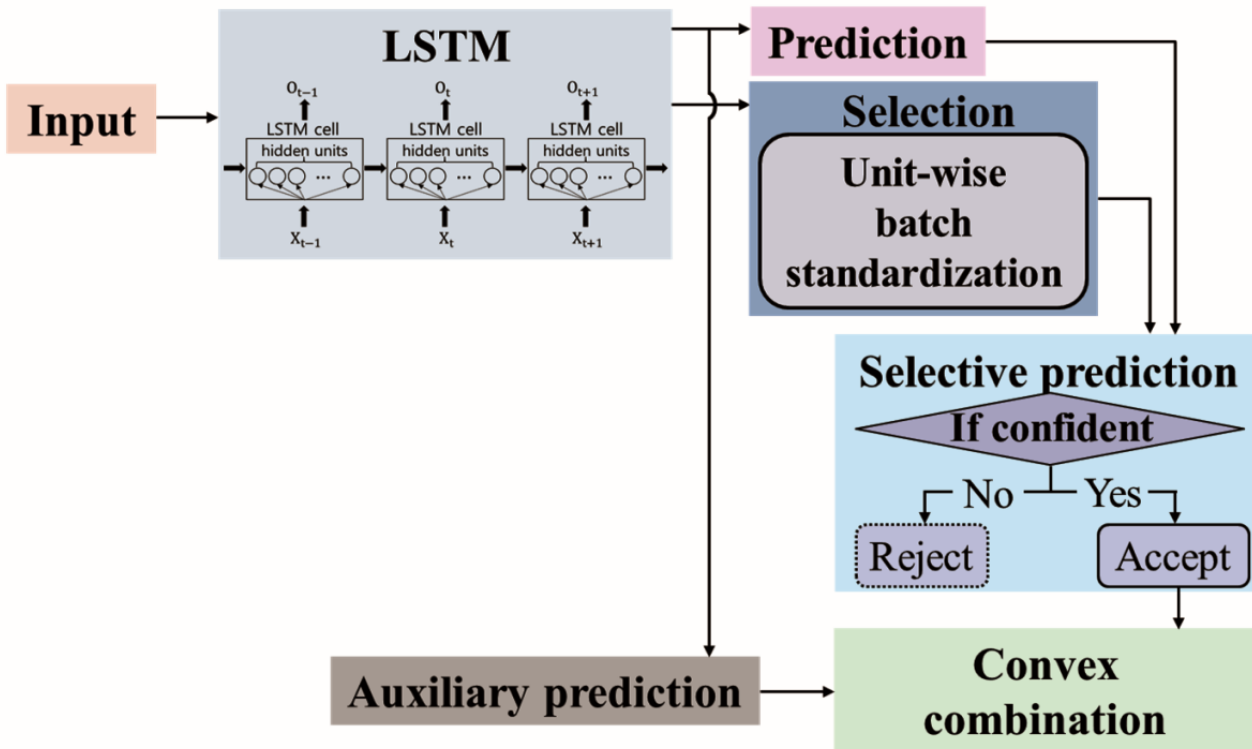
Optimization

An optimization method was used to constrain coverage and reduce the selective risk [31]. The selective prediction model was optimized by the loss functions in equations 6, 7, and 8. This loss function simultaneously regulates the prediction and selection steps. Hence, the selective prediction was regulated to lower the error rate, which is the selective risk for the selected samples according to the prediction confidence. In addition, the selection step was optimized to select training samples based on the predefined target coverage so that the selection step would reject predictions below the confidence level. The target coverage is a controlling hyperparameter for the model to learn the amount of data to be selected during training. On the basis of this, we trained the model so that the coverage value was as close to the target coverage as possible. The target coverage c is in the range $0<c\leq1$. When the parameter set of the selective model (f, g) is Θ the optimization of the selective model is as follows:

The f_θ and g_θ in the selective prediction were optimized by equation 6. It is necessary to constrain coverage and reduce risk (error) for selective prediction. We used the interior point method for optimization [35]. The following unconstrained objective is used to optimize the selective prediction model for a data set S_m:

where c is the target coverage, and λ is a hyperparameter that controls the coverage constraints. Using equation 6, the selection function g is optimized to produce an appropriate prediction confidence score, and the selective prediction is optimized to reduce the selective risk . The empirical coverage value is probabilistically calculated using the selection function. The

Ψ allows the coverage value to approximate the target coverage during the training session. The auxiliary classification loss is optimized using the loss function . Overall, optimization can be defined using a convex combination expressed by the following equations:

where α is another user-controlled parameter for the weights between the selective and auxiliary predictions.

UBS Procedure

In this study, a new selection function structure for LSTM models was designed. The basic frame of the selection function structure was based on a CNN-based model from a previous study [32] that used batch normalization [36] for the selection function. The detailed structure and parameters were determined through a grid search. The output shape of the *many-to-one* structure LSTM is (n_batch, n_hidden_unit), with conventional batch normalization, applying the same mean and variance to all units. However, this method of normalization ignores the features of each hidden unit in the LSTM output. To address this problem, we applied a new UBS that normalizes the batch derived from an original batch normalization [36] while preserving the hidden-unit features captured for each training sample. As shown in Table 1, UBS uses a fully connected layer that maintains the LSTM output's shape while generating the output and standardizing the batch, as shown in Figure 2. When batch normalization is applied to CNNs, normalization factors (mean and variance) are obtained from each input channel [37]. However, to preserve hidden units' individual features, we calculated normalization factors obtained from each LSTM's hidden unit.

Table 1. Detailed structure of the selective prediction step.

Layer	Input shape	Output shape
LSTM[a]	(n_batch, n_time steps, n_features)	(n_batch, n_hidden unit)
FC1[b,c]	(n_batch, n_hidden unit)	(n_batch, n_hidden unit)
FC2[b,d]	(n_batch, n_hidden unit)	(n_batch, n_hidden unit)
ReLU[b,e]	(n_batch, n_hidden unit)	(n_batch, n_hidden unit)
UBS[b,f]	(n_batch, n_hidden unit)	(n_batch, n_hidden unit)
FC3[g]	(n_batch, n_hidden unit)	(n_batch, 1)
Sigmoid	(n_batch, 1)	(n_batch, 1)

[a]LSTM: long short-term memory.

[b]The layer retains the input.

[c]FC1: fully connected layer 1.

[d]FC2: fully connected layer 2.

[e]ReLU: rectified linear unit.

[f]UBS: unit-wise batch standardization.

[g]FC3: fully connected layer 3.

Figure 2. Algorithm of unit-wise batch standardization. LSTM: long short-term memory; ReLU: rectified linear unit.

Algorithm : Unit-wise batch standardization (UBS)

Input : Values of x over after ReLU layer in the selection function O_s :

$$O_s = \begin{bmatrix} x_{11} & \cdots & x_{1n} \\ \vdots & \ddots & \vdots \\ x_{m1} & \cdots & x_{mn} \end{bmatrix}$$

m : mini batch size, n : number of hidden units in LSTM

Scale and shift factors : $\gamma = \{r_{1\ldots n}\}$, $\beta = \{b_{1\ldots n}\}$

Parameters to be learned (tiling scale and shift factors) :

$$\gamma = \begin{bmatrix} r_1 & \cdots & r_n \\ \vdots & \ddots & \vdots \\ r_1 & \cdots & r_n \end{bmatrix}, \ \beta = \begin{bmatrix} b_1 & \cdots & b_n \\ \vdots & \ddots & \vdots \\ b_1 & \cdots & b_n \end{bmatrix}$$

Output : Input standardization along axis 0 :
Unit-wise batch standardization $y = \text{UBS}(O_s)$

for $j = 1$ to n **do**

$\quad \mu_j \leftarrow \frac{1}{m}\sum_{i=1}^{m} x_{ij}$ // mini-batch mean for the jth hidden unit

$\quad \sigma_j^2 \leftarrow \frac{1}{m}\sum_{i=1}^{m}\left(x_{ij} - \mu_j\right)^2$ // mini-batch variance for the jth hidden unit

$\quad$ **for** $i = 1$ to m **do**

$\quad\quad \hat{x}_{ij} \leftarrow \frac{x_{ij} - \mu_j}{\sqrt{\sigma_j^2 + \epsilon}}$ // normalize for the jth hidden unit

$\quad$ **end for**

end for

$\tilde{y} \ \leftarrow \hat{x}$ // Unit-wise batch standardization

$y \ \leftarrow \gamma\tilde{y} + \beta$ // scale and shift

Performance Evaluation

In a health care system, a misdiagnosis involving a type 2 error may imply serious repercussions, and incorrect judgment involving a type 1 error may increase user fatigue. Therefore, we verified the performance of the algorithm by checking false-positive and false-negative rates. The false-positive rate (also known as type 1 error, fall-out, or false-alarm ratio) was calculated as the ratio between the number of negative events incorrectly identified as positive and the total number of actual negative events. The false-negative rate (type 2 error) was calculated as the number of samples misclassified as negative out of the total number of positive events.

Experiment

Overview

Data Sets

This study was reviewed and approved by the institutional review board (#HYUIRB-202111-003) of the Hanyang University, and the requirement for informed consent was waived. A widely used public database was employed to verify

the applicability of the selective prediction model to time series health care data. Considering that the purpose of selective prediction is to reject uncertain predictions, we selected two data sets containing classes that can be misclassified [38-42]: the *human activity recognition using smartphones* and the *Massachusetts Institute of Technology-Beth Israel Hospital* (MIT-BIH) data sets. Detailed descriptions of the data sets have been provided below.

Human Activity Recognition Using Smartphones Data Set

This data set consists of human gait signals monitored by an accelerometer and gyroscope with 6 different activity classes [43]. The signal was measured by attaching Samsung Galaxy S2 smartphones with embedded inertial sensors to the waists of 30 subjects aged 19 to 48 years. Each subject performed six activities (standing, sitting, laying, walking, walking upstairs, and walking downstairs) at least two times for 12 to 15 seconds. The 3-axial linear acceleration and angular velocity were measured at 50 Hz using an embedded accelerometer and gyroscope. The experiments were video-recorded to label the data manually. The signals were preprocessed using a median filter and a third-order low-pass Butterworth filter with a 20-Hz

cutoff frequency and then sampled in sliding windows of 2.56 seconds with 50% overlap (128 readings/window). A total of 10,299 data points were recorded. The training data were randomly selected from 70% of the data set, and the remaining data set was used for the test. The x, y, and z components of the body accelerometer, body gyroscope, and total (gravitational and body) accelerometers were treated as 9 input features. Each sample contained 128 sequences.

MIT-BIH Arrhythmia Data Set

This data set contains 48 half-hour excerpts of two-channel ambulatory electrocardiogram (ECG) recordings from 47 subjects [44]. The recordings were digitized at 360 samples per second per channel with 11-bit resolution over a 10-mV range and annotated independently by 2 or more cardiologists. The data set is publicly available in the PhysioNet [45] database. All protected health information was removed and deidentified using record numbers. A method described in a previous study was used for preprocessing data [46]. First, ECG signals were divided into 10-second intervals. Subsequently, the signal was normalized between 0 and 1. Where the median of the R-R time interval in the ECG signal was T, the time from the R peak to 1.2 T was used as 1 segment. Because the length of the segment changes every 10 seconds, the length of the entire data set is zero-padded based on the longest time. The data set consisted of 109,446 data points with a sampling frequency of 125 Hz. Each data set contained 187 sequences grouped into five classes: N (normal beat), S (supraventricular premature beat), V (premature ventricular contraction), F (fusion of ventricular and normal beats), and Q (unclassifiable beat). Unclassifiable data were not included in this study. As the data for each class were highly imbalanced, 800 data samples were randomly extracted from each class [46]. The data set was sampled for every run, and result was expressed as an average of the results. The data set was then randomly divided into sets: 80% for training and 20% for testing.

Model Architecture and Parameters Setting

Overview

In this study, a selective prediction model was developed using LSTM. Deep learning models such as LSTM are considered effective for extracting meaningful features from raw data. No feature extractor was used in this study because a deep learning model is suitable for use with raw data. The prediction model architecture was determined and optimized based on previous studies, and hyperparameters were optimized using an extensive grid search [47,48]. The details for each data set are described below.

Human Activity Recognition Using Smartphones Data Set

The LSTM model for the human activity recognition using smartphones data set had a single layer with 2 cells and 32

hidden units. For parameter setting, the learning rate was 0.0005, and the L2 regularization was set at a lambda of 0.00005. The mini batch size was 919, and the training epoch was 500. The optimal α and λ were 0.6 and 200, respectively.

MIT-BIH Arrhythmia Data Set

The LSTM model for the MIT-BIH arrhythmia data set had a single layer with 2 cells and 48 hidden units, a learning rate of 0.0001, a minibatch size of 640, and a training epoch of 2000. The optimal α was 0.2, and the optimal λ was 4.

Comparison Method

To prove that the UBS is effective for developing a proper selection function in an LSTM model with a reject option, we compared it with conventional batch normalization and a model without normalization. The false-positive and false-negative rates were also calculated, and a standard LSTM model without a selection function was used as the baseline.

Results

LSTM Performance for Prediction

The baseline models should be optimized for LSTM models without a selection function for each data set. Therefore, we validated the LSTM model prediction performance without any selection. The test accuracies of the LSTM models optimized without a selection step for the human activity recognition using smartphones data set and the MIT-BIH arrhythmia data set are 92.35% and 97.23% for each data set. The precision of the model was 91.72% and the recall was 91.54% for the Human Activity Recognition Using Smartphones data set. For the MIT-BIH arrhythmia data set, the precision of the model was 87.13% and the recall was 78.64%. The F1-score for each data set were 91.63% and 82.67%, respectively.

Coverage Violation

After setting the target coverage, the empirical coverage of the test set was calculated for each normalization method. The target coverage rates were obtained from a previous study [32]. As the target coverage is the target threshold, it should be set to a sufficiently reliable value. Therefore, the target coverages were set at 0.85, 0.90, and 0.95. The difference between the target coverage and the actual coverage value is called *coverage violation*, which estimates the extent to which the model can learn to select the samples as instructed by the target coverage hyperparameter. The experimental results for each data set are listed in Table 2. The coverage value was averaged for 5 different runs. As shown in Table 2, the empirical coverage with UBS produced superior results as they converged on the target coverage, whereas other normalization approaches showed relatively poor results.

Table 2. Empirical coverage of the human activity recognition (HAR) using smartphones and the Massachusetts Institute of Technology-Beth Israel Hospital (MIT-BIH) arrhythmia data sets by different normalization methods. Target coverage was set before training.

Target coverage	HAR using smartphones data set			MIT-BIH arrhythmia data set		
	Normalization method of selective prediction			Normalization method of selective prediction		
	UBS[a]	BN[b]	Without normalization[c]	UBS	BN	Without normalization
0.95, mean (SD)	0.9660 (0.0029)	0.9996 (0.0001)	0.9986 (0.0002)	0.9564 (0.0019)	0.9680 (0.0067)	1.0000 (0)
0.90, mean (SD)	0.9053 (0.0035)	0.9980 (0.0001)	0.9984 (0.0001)	0.9084 (0.0055)	0.9998 (0.0001)	1.0000 (0)
0.85, mean (SD)	0.8582 (0.0007)	0.9237 (0.0026)	0.9986 (0.0002)	0.8888 (0.0016)	0.9518 (0.0001)	1.0000 (0)
Average violation, %	0.98	7.38	9.85	1.79	7.32	10.00

[a]UBS: unit-wise batch standardization.

[b]BN: batch normalization (a normalization method using the mean and variance obtained from the input batch).

[c]*Without normalization* means that there was no normalization in the selection function structure.

Selective Risk (Error Rate)

The selective risks for each normalization method are presented in Table 3. The selective risk value was averaged from 5 different runs. In the selective prediction model with LSTM, the selective risk increased with coverage. UBS normalization achieved relatively superior performance with various target coverages compared with conventional batch normalization. If normalization was not applied, the risk varied widely.

Table 3. Selective risk of the human activity recognition (HAR) using smartphones and the Massachusetts Institute of Technology-Beth Israel Hospital (MIT-BIH) arrhythmia data sets by different normalization methods.

Target coverage	HAR using smartphones data set			MIT-BIH arrhythmia data set		
	Normalization method of selective prediction			Normalization method of selective prediction		
	UBS[a]	BN[b]	Without normalization[c]	UBS	BN	Without normalization
0.95, mean (SD)	0.1423 (0.0041)	0.1611 (0.0445)	0.1476 (0.0068)	0.1970 (0.0038)	0.2175 (0.0108)	0.2000 (0.4472)
0.90, mean (SD)	0.1232 (0.0042)	0.1283 (0.0067)	0.1312 (0.0139)	0.1791 (0.0050)	0.3200 (0.1095)	0.2000 (0.4472)
0.85, mean (SD)	0.1136 (0.0060)	0.1170 (0.0024)	0.1267 (0.0145)	0.1585 (0.0028)	0.1967 (0.0064)	0.2000 (0.4472)
Average risk	0.1264	0.1355	0.1352	0.1782	0.2447	0.2

[a]UBS: unit-wise batch standardization.

[b]BN: batch normalization (a normalization method using the mean and variance obtained from the input batch).

[c]*Without normalization* means that there was no normalization in the selection function structure.

False-Positive and False-Negative Rates

As the selective prediction model produced classification results only when it was confident about its own classification, we expected that both false-positive and false-negative rates would decrease. The false-positive and false-negative rates of each data set were calculated from the results of the model that achieved the best performance among 5 different runs (Tables 4 and 5). The baseline models were well-optimized LSTM models without a selection function for each data set.

Table 4. False-positive rates of the human activity recognition (HAR) using smartphones and the Massachusetts Institute of Technology-Beth Israel Hospital (MIT-BIH) arrhythmia data sets by different normalization methods.

Target coverage	HAR using smartphones data set				MIT-BIH arrhythmia data set			
	Normalization method of selective prediction			General prediction[a]	Normalization method of selective prediction			General prediction
	UBS[b]	BN[c]	Without normalization[d]		UBS	BN	Without normalization	
0.95, %	2.04	2.59	2.65	N/A[e]	6.34	7.67	6.93	N/A
0.90, %	2.00	3.00	2.63	N/A	5.39	6.98	6.77	N/A
0.85, %	2.22	3.02	2.63	N/A	5.66	7.03	7.97	N/A
Average false-positive rate, %	2.09	2.87	2.64	2.89	5.80	7.23	7.22	6.44

[a]General prediction is the long short-term memory classification model's false-positive rate without a selection function.

[b]UBS: unit-wise batch standardization.

[c]BN: batch normalization (a normalization method using the mean and variance obtained from the input batch).

[d]*Without normalization* means that there was no normalization in the selection function structure.

[e]N/A: not applicable.

Table 5. False-negative rates of the human activity recognition (HAR) using smartphones and the Massachusetts Institute of Technology-Beth Israel Hospital (MIT-BIH) arrhythmia data sets by different normalization methods.

Target coverage	HAR using smartphones data set				MIT-BIH arrhythmia data set			
	Normalization method of selective prediction			General prediction[a]	Normalization method of selective prediction			General prediction
	UBS[b]	BN[c]	Without normalization[d]		UBS	BN	Without normalization	
0.95, %	10.18	17.17	12.69	N/A[e]	18.82	23.33	20.78	N/A
0.90, %	10.72	15.04	13.05	N/A	16.48	20.94	20.31	N/A
0.85, %	10.85	14.46	12.94	N/A	16.41	21.44	23.91	N/A
Average false-negative rate, %	10.58	15.56	12.89	14.48	17.24	21.90	21.67	26.47

[a]General prediction is the long short-term memory classification model's false-positive rate without a selection function.

[b]UBS: unit-wise batch standardization.

[c]BN: batch normalization; which is a normalization method using the mean and variance obtained from the input batch.

[d]Without normalization means that there was no normalization in the selection function structure.

[e]N/A: not applicable.

Learned Feature Representation

Figure 3 shows the visualization of the features learned from the LSTM models using t-distributed stochastic neighbor embedding [49]. Figure 3 (left) depicts the test set sample that was not rejected when the target coverage was set at 0.95. The data set used in the visualization was the test set for the human activity recognition using smartphones data set. The *Sitting* (cyan) and *Standing* samples (blue) are more mixed in Figure 3 (right) than in Figure 3 (left). The *Walking_Down_Stairs* (green), *Walking_Up_Stairs* (orange), and *Walking* samples (red) are closely clustered in Figure 3 (left), whereas some of them overlap in Figure 3 (right).

XSL•FO

RenderX

Figure 3. t-Distributed stochastic neighbor embedding visualizations of learned features using all test samples in the human activity recognition using smartphones data set. Left: Long short-term memory with a reject option using unit-wise batch standardization results when the target coverage was 0.95. Rejected samples were not included in this figure. Right: long short-term memory model results without a reject option.
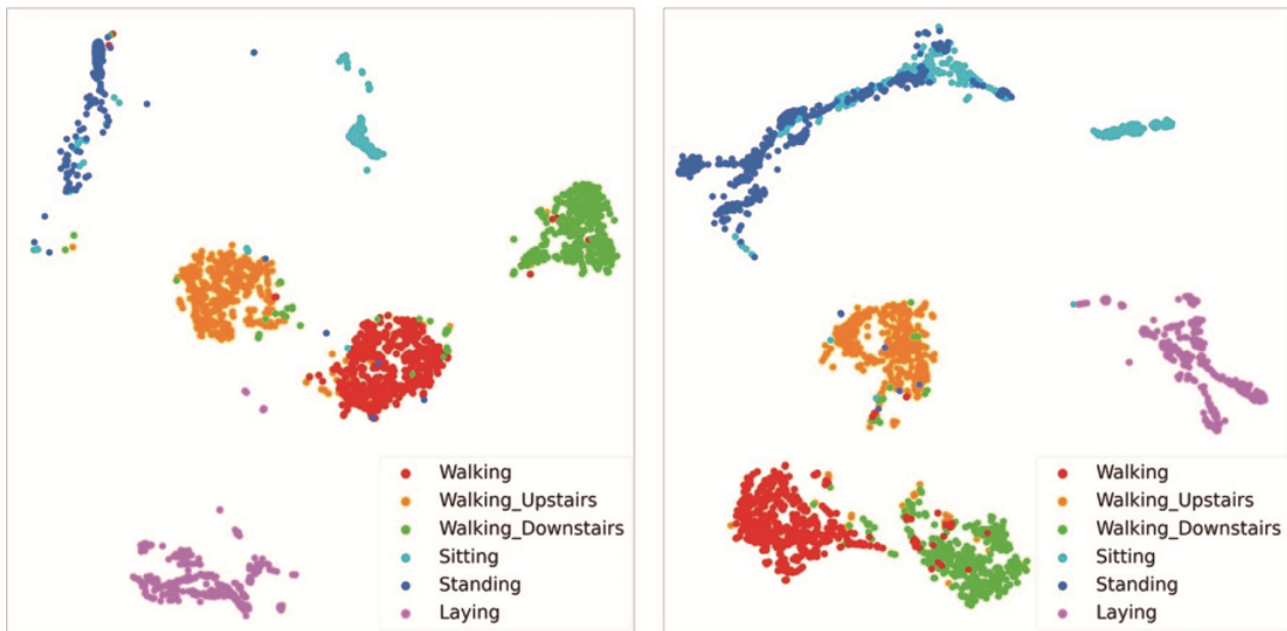
Discussion

Principal Findings

Our objective is to develop a selective prediction model using LSTM. The developed selective prediction model rejected samples using the confidence level of classifications. This selective prediction model with a reject option was trained to determine whether to obtain a classification based on targeted coverage. If the model's classification confidence was low, the model rejected the classification and did not apply information to backpropagate on samples. As a result, the selective prediction model was trained mainly using samples that had a sufficient confidence level, which guaranteed reliability and low error rates for samples that were not rejected. To implement selective prediction for LSTM, we conducted an experiment to identify a method of normalization that could improve the performance of the selection function.

In health care systems, high accuracy is important, but low false-positive and false-negative rates are also essential. To handle various time series data obtained from a health care system, we devised a selective prediction model with LSTM using an effective selection function and focused on the structure of the function. As shown in Table 1, the output of the *many-to-one* LSTM includes hidden-unit information. Our goal was to deal with LSTMs that have *many-to-one* structures, but conventional batch normalization normalizes all batches at once. To tackle this problem, we devised UBS as a special method of normalization that attempts to normalize each hidden unit in LSTM. The false-positive and false-negative rates for each data set were meaningful. For each target coverage, the selective prediction model with UBS was superior to the model with batch normalization and the model without normalization (Tables 4 and 5). These findings show that a selective function using UBS can decrease false-positive and false-negative rates. On this basis, we interpreted that the model with UBS can learn

class-specific features and consider which samples to reject in the training phase.

UBS also helped the model be trained based on target coverage and reduced selective risk. Using 2 public health data sets, the empirical coverage violation of the selective prediction was lower than that of the other 2 methods. The selection function with the UBS had the lowest selective risk (Table 3). The MIT-BIH arrhythmia data set results show that the coverage of the model without normalization was high regardless of the target coverage. These findings imply that the selective function without normalization did not perform as desired. We assumed that these results were based on whether the normalization methods considered hidden-unit characteristics of LSTM.

Regarding the learned feature representation, the classification model with the reject option differed from existing models. In Figure 3, a classification model with the reject option achieved relatively better classification performance than the conventional model without the reject option because the selective prediction LSTM model did not learn the features from samples with a low confidence level. As reported in a previous study [32], this suggests that representational capacity was not wasted because the model was trained mainly on samples with a high confidence level using selective prediction. Using this property, selective prediction allows humans to classify samples with low reliability and act as a second opinion in health care applications. In summary, the selective prediction model successfully classified samples based on high confidence-level features and simultaneously reduced the error rate by using the reject option.

Although our research supports the possibility of generating LSTM models with selective prediction, challenges remain. First, interpretation of the visualization of the learned features is limited in this study and needs to be addressed in further studies. Second, when LSTM was used for selective prediction, it was difficult to optimize parameters that control selection

functions, such as α and λ, for each data set. During the experiments, we used only 2 data sets for testing and targeted only the reject option to determine the confidence level of classifications. In future studies, efficient optimization methods should be devised and applied to various models using various data sets.

Conclusions

In this study, we developed LSTM classification models with a reject option to classify medical data time series. To develop the LSTM classification models with the reject option, UBS was applied. The UBS achieved superior performance (concerning coverage, risk, and false-positive and false-negative rates) compared with 2 other methods of normalization in experiments using 2 public time series data sets.

If the performance in classifying nonrejected samples can be maximized by adjusting coverage or selective risks, humans can trust the output of a highly confident AI model and spend more time on other rejected samples (low confidence). The final performance (human+AI) can be maximized by appropriate automation using selective prediction.

To the best of our knowledge, this is the first study demonstrating the possibility of an LSTM classification model with a reject option for time series data. Our findings may apply to various other time series data sets that require reliability.

Authors' Contributions

This study was originally conceived by BRN. BRN developed a deep learning model and wrote draft of manuscript as a lead author. Data extraction and preprocessing was conducted by BRN and JYK. IYK and BHC jointly supervised this project as co-corresponding authors. All authors provided critical feedback and approved the final manuscript.

Conflicts of Interest

None declared.

References

1. Rong G, Mendez A, Bou Assi E, Zhao B, Sawan M. Artificial intelligence in healthcare: review and prediction case studies. Engineering 2020;6(3):291-301. [doi: 10.1016/j.eng.2019.08.015]
2. Moon JH, Lee DY, Cha WC, Chung MJ, Lee KS, Cho BH, et al. Automatic stenosis recognition from coronary angiography using convolutional neural networks. Comput Methods Programs Biomed 2021;198:105819 [FREE Full text] [doi: 10.1016/j.cmpb.2020.105819] [Medline: 33213972]
3. Kim JY, Ro K, You S, Nam BR, Yook S, Park HS, et al. Development of an automatic muscle atrophy measuring algorithm to calculate the ratio of supraspinatus in supraspinous fossa using deep learning. Comput Methods Programs Biomed 2019;182:105063. [doi: 10.1016/j.cmpb.2019.105063] [Medline: 31505380]
4. Xue Q, Chuah MC. Explainable deep learning based medical diagnostic system. Smart Health 2019;13:100068. [doi: 10.1016/j.smhl.2019.03.002]
5. Colak E, Moreland R, Ghassemi M. Five principles for the intelligent use of AI in medical imaging. Intensive Care Med 2021;47(2):154-156. [doi: 10.1007/s00134-020-06316-8] [Medline: 33449134]
6. Quinn TP, Senadeera M, Jacobs S, Coghlan S, Le V. Trust and medical AI: the challenges we face and the expertise needed to overcome them. J Am Med Inform Assoc 2021;28(4):890-894 [FREE Full text] [doi: 10.1093/jamia/ocaa268] [Medline: 33340404]
7. Hengstler M, Enkel E, Duelli S. Applied artificial intelligence and trust—the case of autonomous vehicles and medical assistance devices. Technol Forecast Soc Change 2016;105:105-120. [doi: 10.1016/j.techfore.2015.12.014]
8. Chow CK. An optimum character recognition system using decision functions. IRE Trans Electron Comput 1957;EC-6(4):247-254. [doi: 10.1109/tec.1957.5222035]
9. Wang J, Chen Y, Hao S, Peng X, Hu L. Deep learning for sensor-based activity recognition: a survey. Pattern Recognit Lett 2019;119:3-11. [doi: 10.1016/j.patrec.2018.02.010]
10. Miotto R, Wang F, Wang S, Jiang X, Dudley JT. Deep learning for healthcare: review, opportunities and challenges. Brief Bioinform 2018;19(6):1236-1246 [FREE Full text] [doi: 10.1093/bib/bbx044] [Medline: 28481991]
11. Lipton ZC, Kale DC, Elkan C, Wetzel RC. Learning to diagnose with LSTM recurrent neural networks. In: Proceedings of the 4th International Conference on Learning Representations. 2016 Presented at: ICLR '16; May 2-4, 2016; San Juan, Puerto Rico. [doi: 10.1093/acref/9780195301731.013.43262]
12. Choi E, Schuetz A, Stewart WF, Sun J. Using recurrent neural network models for early detection of heart failure onset. J Am Med Inform Assoc 2017;24(2):361-370 [FREE Full text] [doi: 10.1093/jamia/ocw112] [Medline: 27521897]

13. Razavian N, Marcus J, Sontag D. Multi-task prediction of disease onsets from longitudinal laboratory tests. In: Proceedings of the 1st Machine Learning for Healthcare Conference. 2016 Presented at: PMLR '16; August 19-20, 2016; Los Angeles, CA p. 73-100.

14. Reddy BK, Delen D. Predicting hospital readmission for lupus patients: an RNN-LSTM-based deep-learning methodology. Comput Biol Med 2018;101:199-209. [doi: 10.1016/j.compbiomed.2018.08.029] [Medline: 30195164]

15. Şentürk Ü, Yücedağ I, Polat K. Repetitive neural network (RNN) based blood pressure estimation using PPG and ECG signals. In: 2nd International Symposium on Multidisciplinary Studies and Innovative Technologies. 2018 Presented at: ISMSIT '18; October 19-21, 2018; Ankara, Turkey p. 1-4. [doi: 10.1109/ismsit.2018.8567071]

16. Su P, Ding XR, Zhang YT, Liu J, Miao F, Zhao N. Long-term blood pressure prediction with deep recurrent neural networks. In: IEEE EMBS International Conference on Biomedical & Health Informatics. 2018 Presented at: BHI '18; March 4-7, 2018; Las Vegas, NV p. 323-328. [doi: 10.1109/bhi.2018.8333434]

17. Xu X, Jeong S, Li J. Interpretation of electrocardiogram (ECG) rhythm by combined CNN and BiLSTM. IEEE Access 2020;8:125380-125388. [doi: 10.1109/access.2020.3006707]

18. Rana A, Kim KK. ECG heartbeat classification using a single layer LSTM model. In: International SoC Design Conference. 2019 Presented at: ISOCC '19; October 6-9, 2019; Jeju, South Korea p. 267-268. [doi: 10.1109/isocc47750.2019.9027740]

19. Hernández F, Suárez LF, Villamizar J, Altuve M. Human activity recognition on smartphones using a bidirectional LSTM network. In: XXII Symposium on Image, Signal Processing and Artificial Vision. 2019 Presented at: STSIVA '19; April 24-26, 2019; Bucaramanga, Colombia p. 1-5. [doi: 10.1109/stsiva.2019.8730249]

20. Hammerla NY, Halloran S, Plötz T. Deep, convolutional, and recurrent models for human activity recognition using wearables. In: Proceedings of the 25th International Joint Conference on Artificial Intelligence. 2016 Presented at: IJCAI '16; July 9-15, 2016; New York, NY p. 1533.

21. Chowdhury SS, Hasan MS, Sharmin R. Robust heart rate estimation from PPG signals with intense motion artifacts using cascade of adaptive filter and recurrent neural network. In: 2019 IEEE Region 10 Conference. 2019 Presented at: TENCON '19; October 17-20, 2019; Kochi, India p. 1952-1957. [doi: 10.1109/tencon.2019.8929692]

22. Zhao Y, Yang R, Chevalier G, Xu X, Zhang Z. Deep residual Bidir-LSTM for human activity recognition using wearable sensors. Math Probl Eng 2018;2018:1-13. [doi: 10.1155/2018/7316954]

23. Cordella LP, De Stefano C, Tortorella F, Vento M. A method for improving classification reliability of multilayer perceptrons. IEEE Trans Neural Netw 1995;6(5):1140-1147. [doi: 10.1109/72.410358] [Medline: 18263404]

24. De Stefano C, Sansone C, Vento M. To reject or not to reject: that is the question-an answer in case of neural classifiers. IEEE Trans Syst, Man, Cybern C 2000;30(1):84-94. [doi: 10.1109/5326.827457]

25. El-Yaniv R, Wiener Y. Pointwise tracking the optimal regression function. In: Advances in Neural Information Processing Systems 25. 2012 Presented at: NIPS '12; December 3-8, 2012; Lake Tahoe, NV p. 2042-2050.

26. Geifman Y, El-Yaniv R. Selective classification for deep neural networks. In: Proceedings of the 31st International Conference on Neural Information Processing Systems. 2017 Presented at: NIPS '17; December 4-9, 2017; Long Beach, CA p. 4885-4894. [doi: 10.7551/mitpress/11474.003.0014]

27. Blundell C, Cornebise J, Kavukcuoglu K, Wierstra D. Weight uncertainty in neural network. In: Proceedings of The 32nd International Conference on Machine Learning. 2015 Presented at: ICML '15; July 6-11, 2015; Lille, France.

28. Lipton Z, Li X, Gao J, Li L, Ahmed F, Deng L. BBQ-networks: efficient exploration in deep reinforcement learning for task-oriented dialogue systems. In: Proceedings of the 32nd AAAI Conference on Artificial Intelligence. 2018 Presented at: AAAI '18; February 2-7, 2018; New Orleans, LA.

29. Houthooft R, Chen X, Duan Y, Schulman J, De Turck F, Abbeel P. VIME: variational information maximizing exploration. In: Proceedings of the 30th International Conference on Neural Information Processing Systems. 2016 Presented at: NIPS '16; December 5-10, 2016; Barcelona, Spain p. 1117-1125. [doi: 10.1016/S0377-0427(00)00433-7]

30. Fortunato M, Blundell C, Vinyals O. Bayesian recurrent neural networks. arXiv (forthcoming) 2017:1-14 [FREE Full text]

31. Cortes C, DeSalvo G, Mohri M. Learning with rejection. In: Proceedings of the 27th International Conference on Algorithmic Learning Theory. 2016 Presented at: ALT '016; October 19-21, 2016; Bari, Italy p. 67-82. [doi: 10.1007/978-3-319-46379-7_5]

32. Geifman Y, El-Yaniv R. SelectiveNet: a deep neural network with an integrated reject option. In: Proceedings of The 36th International Conference on Machine Learning. 2019 Presented at: ICML '19; June 10-15, 2019; Long Beach, CA.

33. Hochreiter S, Schmidhuber J. Long short-term memory. Neural Comput 1997;9(8):1735-1780. [doi: 10.1162/neco.1997.9.8.1735] [Medline: 9377276]

34. El-Yaniv R, Wiener Y. On the foundations of noise-free selective classification. J Mach Learn Res 2010;11(53):1605-1641.

35. Potra FA, Wright SJ. Interior-point methods. J Comput Appl Math 2000;124(1-2):281-302. [doi: 10.1016/s0377-0427(00)00433-7]

36. Ioffe S, Szegedy C. Batch normalization: accelerating deep network training by reducing internal covariate shift. In: Proceedings of the 32nd International Conference on Machine Learning. 2015 Presented at: ICML '15; July 6-11, 2015; Lille, France p. 448-456.

37. Wu Y, He K. Group normalization. In: Proceedings of the 15th European Conference on Computer Vision. 2018 Presented at: ECCV '18; September 8-14, 2018; Munich, Germany p. 3-19. [doi: 10.1007/978-3-030-01261-8_1]

38. Bulbul E, Cetin A, Dogru IA. Human activity recognition using smartphones. In: Proceedings of the 2nd International Symposium on Multidisciplinary Studies and Innovative Technologies. 2018 Presented at: ISMSIT '18; October 19-21, 2018; Ankara, Turkey p. 1-6. [doi: 10.1109/ismsit.2018.8567275]

39. Wan S, Qi L, Xu X, Tong C, Gu Z. Deep learning models for real-time human activity recognition with smartphones. Mobile Netw Appl 2019;25(2):743-755. [doi: 10.1007/s11036-019-01445-x]

40. Xia K, Huang J, Wang H. LSTM-CNN architecture for human activity recognition. IEEE Access 2020;8:56855-56866. [doi: 10.1109/access.2020.2982225]

41. Ge Z, Zhu Z, Feng P, Zhang S, Wang J, Zhou B. ECG-signal classification using SVM with multi-feature. In: The 8th IEEE International Symposium on Next-Generation Electronics. 2019 Presented at: ISNE '19; October 9-10, 2019; Zhengzhou, China. [doi: 10.1109/isne.2019.8896430]

42. Desai U, Martis RJ, Nayak CG, Sarika K, Seshikala G. Machine intelligent diagnosis of ECG for arrhythmia classification using DWT, ICA and SVM techniques. In: 2015 Annual IEEE India Conference. 2015 Presented at: INDICON '15; December 17-20, 2015; New Delhi, India. [doi: 10.1109/indicon.2015.7443220]

43. Anguita D, Ghio A, Oneto L, Parra X, Reyes-Ortiz JL. A public domain dataset for human activity recognition using smartphones. In: Proceedings of 2013 European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. 2013 Presented at: ESANN '13; April 24-26, 2013; Bruges, Belgium p. 437-442.

44. Moody GB, Mark RG. The impact of the MIT-BIH arrhythmia database. IEEE Eng Med Biol Mag 2001;20(3):45-50. [doi: 10.1109/51.932724] [Medline: 11446209]

45. Goldberger AL, Amaral LA, Glass L, Hausdorff JM, Ivanov PC, Mark RG, et al. PhysioBank, PhysioToolkit, and PhysioNet: components of a new research resource for complex physiologic signals. Circulation 2000;101(23):E215-E220. [doi: 10.1161/01.cir.101.23.e215] [Medline: 10851218]

46. Kachuee M, Fazeli S, Sarrafzadeh M. ECG heartbeat classification: a deep transferable representation. In: 2018 IEEE International Conference on Healthcare Informatics. 2018 Presented at: ICHI '18; June 4-7, 2018; New York, NY p. 443-444. [doi: 10.1109/ichi.2018.00092]

47. Meng L, Zhao B, Chang B, Huang G, Sun W, Tung F, et al. Interpretable spatio-temporal attention for video action recognition. In: 2019 IEEE/CVF International Conference on Computer Vision Workshop. 2019 Presented at: ICCVW '19; October 27-28, 2019; Seoul, South Korea p. 1513-1522. [doi: 10.1109/iccvw.2019.00189]

48. Zhang J, Fan DP, Dai Y, Anwar S, Saleh FS, Zhang T, et al. UC-Net: uncertainty inspired RGB-D saliency detection via conditional variational autoencoders. In: 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2020 Presented at: CVPR '20; June 13-19, 2020; Seattle, WA p. 8579-8588. [doi: 10.1109/cvpr42600.2020.00861]

49. Van der Maaten L, Hinton G. Visualizing data using t-SNE. J Mach Learn Res 2008;9(86):2579-2605.

Abbreviations

AI: artificial intelligence
CNN: convolutional neural network
ECG: electrocardiogram
LSTM: long short-term memory
MC: Monte Carlo
MIT-BIH: Massachusetts Institute of Technology-Beth Israel Hospital
RNN: recurrent neural network
UBS: unit-wise batch standardization

XSL•FO
RenderX

Original Paper

Vascular Aging Estimation Based on Artificial Neural Network Using Photoplethysmogram Waveform Decomposition: Retrospective Cohort Study

Junyung Park[1], MSc; Hangsik Shin[2], PhD

[1]Department of Biomedical Engineering, Chonnam National University, Yeosu, Republic of Korea

[2]Department of Convergence Medicine, Asan Medical Center, University of Ulsan College of Medicine, Seoul, Republic of Korea

Corresponding Author:
Hangsik Shin, PhD
Department of Convergence Medicine
Asan Medical Center
University of Ulsan College of Medicine
388-1 Pungnap-dong
Songpa-gu
Seoul, 05505
Republic of Korea
Phone: 82 2 3010 2099
Email: hangsik.shin@gmail.com

Abstract

Background: For the noninvasive assessment of arterial stiffness, a well-known indicator of arterial aging, various features based on the photoplethysmogram and regression methods have been proposed. However, whether because of the existing characteristics not accurately reflecting the characteristics of the incident and reflected waveforms of the photoplethysmogram or because of the lack of expressive power of the regression model, a reliable arterial stiffness assessment technique based on a single photoplethysmogram has not yet been proposed.

Objective: The purpose of this study is to discover highly correlated features from the incident and reflected waves decomposed from a photoplethysmogram waveform and to develop an artificial neural network-based regression model for the assessment of vascular aging using newly derived features.

Methods: We obtained photoplethysmograms from 757 participants. All recorded photoplethysmograms were segmented for each beat, and each waveform was decomposed into incident and reflected waves by the Gaussian mixture model. The 26 basic features and 52 combined features were defined from the morphological characteristics of the incident and reflected waves. The regression model of the artificial neural network was developed using the defined features.

Results: In correlation analysis, the features from the amplitude of the reflected wave and the skewness of the photoplethysmogram showed a relatively strong correlation with the participant's real age. In the estimation of real age, the artificial neural network model showed 10.0 years of root mean square error. Its estimated age and real age had a strong correlation of 0.63 ($P<.001$).

Conclusions: This study proved that the features defined from the reflected wave and skewness of the photoplethysmogram are useful to assess vascular aging. Moreover, the regression model of artificial neural network using these features shows the feasibility for the estimation of vascular aging.

Introduction

Arterial stiffness is one of the major factors to clinically assess the risk of cardiovascular disease [1]. Hemodynamically, it is known that arterial stiffness increases with aging because of the change of arterial composition and the reduction of arterial elasticity [2]. Therefore, it is possible to objectively grasp the aging status of arteries through arterial stiffness. An increase in arterial stiffness indicates the aging of blood vessels, while

a decrease in arterial stiffness indicates the health of blood vessels [3,4]. In previous studies, the assessment of arterial stiffness was conducted with features or blood pressure values extracted from continuous blood pressure waveforms [5-9]. Murgo et al [5] observed continuous changes in arterial stiffness with age using the augmentation index (AIx), which is defined as the percentage of central augmented pressure to central pulse pressure of the blood pressure waveform. According to a study by McEniery et al [7], it is possible to accurately measure arterial stiffness using the AIx calculated from aortic blood pressure waveforms, but it is reported that AIx is a sensitive marker only for those under 50 years of age. In the assessment of arterial stiffness with AIx, continuous blood pressure waveforms must be measured in an invasive way. Thus, it puts a burden on the patients and is difficult to measure in daily life. Antza et al [9] classified the presence or absence of early vascular aging from the blood pressure data using the machine learning method of random forest. However, Antza et al [9] only determined the presence or absence of vascular aging but could not explain the continuous process of vascular aging.

Photoplethysmogram (PPG), which is a noninvasive optical measuring technique of blood volume changes in microvessels, was also used to assess arterial stiffness. In the PPG waveform, the systolic phase and the diastolic phase repeatedly appear, corresponding to the cardiac systole and the cardiac diastole. The systolic phase indicates an increase in vascular blood volume, and the diastolic phase indicates a decrease in vascular blood volume [10]. Millasseau et al [11,12] addressed that the PPG waveform is formed by the superposition of the incident wave and the reflected wave of the blood pressure. The incident wave is generated by cardiac systole, and the reflected wave is generated by impedance mismatch at arterial bifurcation points. Dawber et al [13] also analyzed how the shape of PPG waveform changes according to the increase in arterial stiffness due to aging. They found that as aging progresses, the diacritic notch of the PPG waveform gradually disappears and the returning time of the reflected wave is shortened. Therefore, their study showed that changes of the PPG waveform could be used to evaluate arterial stiffness. Millasseau et al [14,15] derived the stiffness index (SI) based on the time difference between the systolic and diastolic peaks of PPG, and reported that SI has a significant difference according to vascular aging. Further, Yousef et al [16] calculated the reflection index (RI) as the ratio of PPG's systolic amplitude to diastolic amplitude and showed that RI significantly increased with age. However, the RI and SI introduced in both studies are obtained from the summed waveform of the incident and reflected waves of the PPG, despite the concept being derived from the individual incident and reflected waves of the PPG. Therefore, these features cannot be said to accurately reflect the incident and reflected wave characteristics of the PPG and may be influenced by other external factors. Park et al [17] used the wave decomposition method to define features and develop the vascular assessment model in their study. They decomposed a PPG waveform into an incident wave and a reflected wave and defined features from the waves, directly reflecting the incident and reflected characteristics of PPG. They then confirmed that the defined features had a higher correlation with age than RI and SI and developed a regression model for vascular aging assessment.

In recent studies, machine learning techniques have been introduced to evaluate arterial stiffness. Dall'Olio et al [18] created a convolutional neural network (CNN)-based vascular aging assessment model, which used the PPG raw signal measured by smartphone as an input. Their CNN-based model showed similar performance to the existing PPG feature-based model, and it verified that the machine learning models have the possibility of vascular aging assessment with input data measured from a wearable device. Chiarelli et al [19] estimated the actual age of participants from PPG and electrocardiogram (ECG) measurement, using a deep convolutional neural network (DCNN) model. Their DCNN model showed the result of 7-year-old root mean squared error (RMSE), which has a higher performance in vascular aging estimation than the PPG-feature-based multiple regression and artificial neural network (ANN) models.

The purpose of this study is to develop a new vascular aging assessment model using the PPG, which could be noninvasively and easily measured in daily life. In particular, unlike the existing PPG-based vascular aging estimation studies, we decompose the incident and reflected waves of the PPG waveform. New highly correlated features are then explored for vascular aging assessment from the decomposed PPG waves. Lastly, an ANN-based regression model with excellent nonlinear estimation performance is applied to estimate vascular aging.

Methods

Data and Ethical Considerations

Data were obtained from a total of 1000 patients who were scheduled for elective surgery (thyroid, breast, or abdominal) from July to September 2015 at Asan Medical Center. Through cross-checking of two researchers, 17 participants with loss of signal and 226 participants with indistinguishable PPG waveforms were excluded from the analysis. As a result, data from a total of 757 participants were used. Table 1 shows the summarized characteristics of 757 participants included in the analysis. The PPG waveform was obtained using a pulse oximeter (E^2-KIT; KTMED, Co Ltd), and the PPG Probe was placed between the nasal column and the nasal septum as a transmit type. Signals were recorded at 125 or 250 Hz sampling frequency for 5 min. Data acquisition was performed after obtaining approval from the Asan Medical Center (Songpa-gu, Seoul, South Korea) Research Ethics Committee (IRB No.2015-0104).

Table 1. Characteristics of patients included in the analysis (N=757).

Category	Values
Sex, n (%)	
Male	348 (46.0)
Female	409 (54.0)
ASA PS[a], n (%)	
PS 1	450 (59.4)
PS 2	277 (36.6)
PS 3	30 (4.0)
Weight (kg), median (range)	61.8 (54.1-69.4)
Height (cm), median (range)	161.6 (155.7-168.0)
BMI (kg/m^2), median (range)	23.5 (21.3-25.9)
Age (years)[b], n (%)	
0-29	10 (1.3)
30-39	61 (8.1)
40-49	168 (22.2)
50-59	215 (28.4)
60-69	177 (23.4)
70-79	108 (14.3)
80-89	18 (2.4)
Social characteristics	
Smoking	111 (14.7)
Alcohol	240 (31.7)
Medical history (multiple answers possible) , n (%)	
Hypertension	213 (28.1)
Diabetes mellitus	90 (11.9)
Pulmonary disease[c]	15 (2.0)
Renal disease[d]	5 (0.7)
Hepatic disease[e]	23 (3.0)
Neurologic disease[f]	8 (1.1)
Others[g]	16 (2.1)

[a]ASA PS: American Society of Anesthesiologists Physical Status((1) a normal healthy patient, (2) a patient with mild systemic disease, and (3) a patient with severe systemic disease).

[b]The median age is 56 years, with a range of 46-65 years.

[c]Pulmonary disease: asthma (7), emphysema (1), bronchiectasis (1), chronic obstructive pulmonary disease (5), and old tuberculosis (1).

[d]Renal disease: chronic kidney disease (2) and end stage renal disease (3).

[e]Hepatic disease: hepatitis B virus (11), hepatitis C virus (2), and liver cirrhosis (10).

[f]Neurologic disease: stroke (1) and cardiovascular accident (7).

[g]Others: angina (12), carotid artery stenosis (1), iron deficiency anemia (1), hyponatremia (1), and intracranial hemorrhage (1).

Preprocessing

The measured signal was filtered using a finite impulse response bandpass filter having a 0.5-10 Hz passband, and then the pulse onset (ie, the start point of the waveform for each pulse) was detected (Figure 1). Based on the detected pulse onset, each participant's PPG was divided into pulses to generate segments. At that time, an arrhythmic waveform with an irregular PPG interval or amplitude, or an abnormal waveform with a maximum diastolic amplitude ($DIA_{peak-amp}$) greater than the systolic maximum amplitude ($SYS_{peak-amp}$), was excluded from the analysis. Since the number of samples for each segment was

different due to the nonuniform heartbeat interval for each participant, linear interpolation was performed so that each segment had the same number of samples (ie, 1000). Since the PPG amplitude measured from each participant has an arbitrary value, it was converted to the value between 0 and 1 using the min-max normalization method.

Figure 1. Characteristics of the original PPG, incident and reflected waves, and reconstructed PPG for deriving candidate features. DIA: diastolic; INC: incident wave; INF: inflection point; OPPG: original photoplethysmogram; PPG: photoplethysmogram; REF: reflected wave; RPPG: reconstructed photoplethysmogram; SYS: systolic.
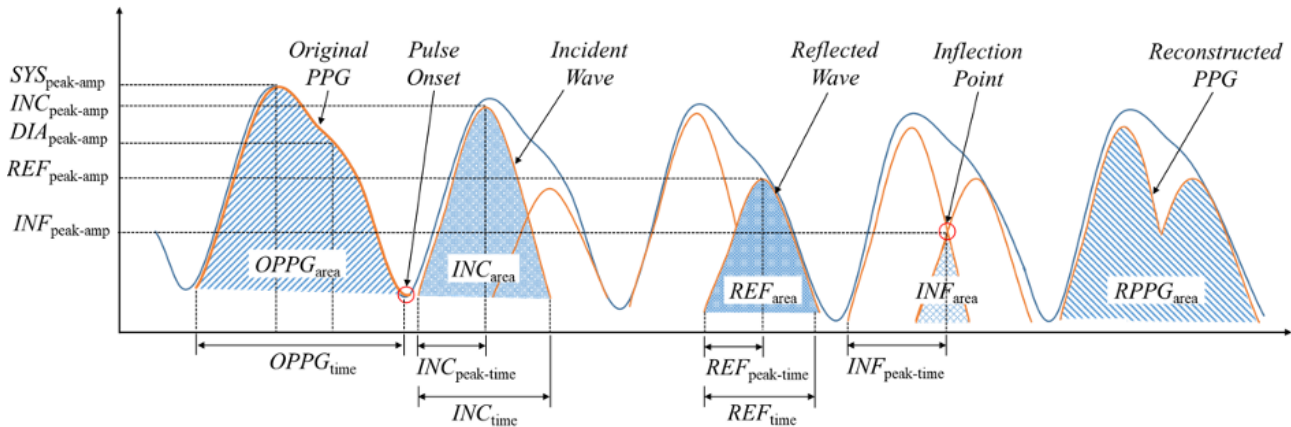
Features

The features for vascular aging assessment consist of a basic feature defined from the specific points of the waveform before and after the decomposition of the incident and reflected waves of the PPG and a combined feature generated by combining the basic feature. Gaussian mixture model [20] was used for PPG waveform decomposition. Figure 1 shows that through waveform decomposition, each PPG segment was divided into two partial waveforms, one incident wave, and one reflected wave. The evaluation of the appropriateness of the PPG waveform decomposition was performed by calculating the correlation coefficient between the reconstructed PPG and the original PPG and comparing the decomposed waveform feature points. In the verification process, only those segments in which the correlation coefficient between the PPG waveform reconstructed from the decomposed incident and reflected waves and the original PPG was 0.9 or more, and the amplitude of the peak of the incident wave ($INC_{peak-amp}$) was greater than the amplitude of the peak of the reflected wave ($REF_{peak-amp}$), were used for analysis.

From the waveforms before and after the decomposition of the incident and reflected waves of the PPG, 26 basic features were generated for the development of the vascular aging estimation model. Table 2 shows the features that are defined as follows: 12 features from the amplitude and time of the maximum peak, and total area, total time, skewness, and kurtosis in each waveform of the incident and reflected waves; 3 features from the amplitude, time, and area under the inflection point where the incident wave and the reflected wave intersect; and 3 features from the area, skewness, and kurtosis of the PPG reconstructed by combining the incident and reflected waves. In addition, 8 features were defined from the feature points indicating the amplitude and time of the systolic and diastolic peaks, the total area and time, and the skewness and kurtosis of the original PPG before the decomposition of the incident and reflected waves. Textbox 1 shows 52 combined features, which were defined as ratios or differences of the 26 basic features after dividing them into time-related features and amplitude-related features. A total of 78 candidate features were generated to develop a regression model for ANN-based vascular aging estimation. All preprocessing and feature extraction processes were performed using Matlab 2018b (Mathworks).

Table 2. Basic features defined from incident and reflected waves, first inflection point, reconstructed PPG[a], and original PPG.

Pulse type and feature	Definition
Incident wave	
$INC^b_{peak-amp}$	Amplitude of incident wave's peak
INC_{area}	Area of incident wave
$INC_{peak-time}$	Time of incident wave's peak
INC_{time}	Time period of incident wave
INC_{skew}	Skewness of incident wave
INC_{kurt}	Kurtosis of incident wave
Reflected wave	
$REF^c_{peak-amp}$	Amplitude of reflected wave's peak
REF_{area}	Area of reflected wave
$REF_{peak-time}$	Time of reflected wave's peak
REF_{time}	Time period of reflected wave
REF_{skew}	Skewness of reflected wave
REF_{kurt}	Kurtosis of reflected wave
First inflection point	
$INF^d_{peak-amp}$	Amplitude of first inflection point
$INF_{peak-time}$	Time of first inflection point
INF_{area}	Area of first inflection
Reconstructed PPG	
$RPPG^e_{area}$	Area of reconstructed PPG
$RPPG_{skew}$	Skewness of reconstructed PPG
$RPPG_{kurt}$	Kurtosis of reconstructed PPG
Original PPG	
$SYS^f_{peak-amp}$	Amplitude of systolic peak
$SYS_{peak-time}$	Time of systolic peak
$DIA^g_{peak-amp}$	Amplitude of diastolic peak
$DIA_{peak-time}$	Time of diastolic peak
$OPPG^h_{area}$	Area of original PPG
$OPPG_{time}$	Time period of original PPG
$OPPG_{skew}$	Skewness of original PPG
$OPPG_{kurt}$	Kurtosis of original PPG

[a]PPG: photoplethysmogram.

[b]INC: incident wave.

[c]REF: reflected wave.

[d]INF: inflection point.

[e]RPPG: reconstructed photoplethysmogram.

[f]SYS: systolic.

[g]DIA: diastolic.

[h]OPPG: original photoplethysmogram.

Textbox 1. Combined features derived from the basic features in the spatial, temporal, and spatiotemporal domains. INC: incident wave; REF: reflected wave; RPPG: reconstructed photoplethysmogram; SYS: systolic; OPPG: original photoplethysmogram.

Domain and feature

Spatial

- $INC^a_{peak\text{-}amp}/INC_{area}$

- $INC_{peak\text{-}amp}/REF^b_{peak\text{-}amp}$

- $INC_{peak\text{-}amp}/REF_{area}$

- $INC_{peak\text{-}amp}/RPPG^c_{area}$

- $INC_{area}/REF_{peak\text{-}amp}$

- INC_{area}/REF_{area}

- $INC_{area}/RPPG_{area}$

- $REF_{peak\text{-}amp}/REF_{area}$

- $REF_{peak\text{-}amp}/RPPG_{area}$

- $REF_{area}/RPPG_{area}$

- $INC_{peak\text{-}amp}-REF_{peak\text{-}amp}$

- $INC_{area}-REF_{area}$

- $SYS^d_{peak\text{-}amp}-INC_{peak\text{-}amp}$

- $SYS_{peak\text{-}amp}-REF_{peak\text{-}amp}$

Temporal

- $INC_{peak\text{-}time}/INC_{time}$

- $INC_{peak\text{-}time}/REF_{peak\text{-}time}$

- $INC_{peak\text{-}time}/REF_{time}$

- $INC_{peak\text{-}time}/OPPG^e_{time}$

- $INC_{time}/REF_{peak\text{-}time}$

- INC_{time}/REF_{time}

- $INC_{time}/OPPG_{time}$

- $REF_{peak\text{-}time}/REF_{time}$

- $REF_{peak\text{-}time}/OPPG_{time}$

- $REF_{time}/OPPG_{time}$

- $REF_{peak\text{-}time}-INC_{peak\text{-}time}$

- $OPPG_{time}-INC_{peak\text{-}time}$

- $OPPG_{time}-REF_{peak\text{-}time}$

Spatiotemporal

- $INC_{peak\text{-}amp}/INC_{peak\text{-}time}$

- $INC_{peak\text{-}amp}/INC_{time}$

- $INC_{peak\text{-}amp}/REF_{peak\text{-}time}$

- $INC_{peak\text{-}amp}/REF_{time}$

- $INC_{peak\text{-}amp}/OPPG_{time}$

- $INC_{area}/INC_{peak\text{-}time}$

- $INC_\text{area}/INC_\text{time}$
- $INC_\text{area}/REF_\text{peak-time}$
- $INC_\text{area}/REF_\text{time}$
- $INC_\text{area}/OPPG_\text{time}$
- $REF_\text{peak-amp}/INC_\text{peak-time}$
- $REF_\text{peak-amp}/INC_\text{time}$
- $REF_\text{peak-amp}/REF_\text{peak-time}$
- $REF_\text{peak-amp}/REF_\text{time}$
- $REF_\text{peak-amp}/OPPG_\text{time}$
- $REF_\text{area}/INC_\text{peak-time}$
- $REF_\text{area}/INC_\text{time}$
- $REF_\text{area}/REF_\text{peak-time}$
- $REF_\text{area}/REF_\text{time}$
- $REF_\text{area}/OPPG_\text{time}$
- $RPPG_\text{area}/INC_\text{peak-time}$
- $RPPG_\text{area}/INC_\text{time}$
- $RPPG_\text{area}/REF_\text{peak-time}$
- $RPPG_\text{area}/REF_\text{time}$
- $RPPG_\text{area}/OPPG_\text{time}$

Artificial Neural Network Regression Model

In this study, since the actual age of participants is estimated based on various features extracted from their PPG, we used the ANN model, which is frequently used for nonlinear regression with independent features. Table 3 shows that an ANN-based regression model for estimating vascular aging was developed and evaluated using the parameters of various conditions. As a result, the model showing the optimal performance was found as indicated in bold. Figure 2 shows that the developed ANN-based regression model consists of an input layer, a hidden layer, and an output layer. The input layer consists of 78 nodes that receive the features defined by the PPG as inputs. The hidden layer consists of a single layer with 128 nodes. The output layer consists of a single node that outputs the age of the participants estimated through calculation in the hidden layer. Rectified linear unit was used as the activation function [21]. Dropout, which removes hidden layer nodes at a certain rate, was applied with the dropout rate of 0.5.

Adam optimizer and learning rate of 0.001 were applied to train the model.

A leave-one-out cross-validation (LOOCV) was used for the development and testing of the ANN-based regression model. In LOOCV, the entire data was divided into one test set and the rest assigned to the model development set. The model development set was divided into a training set and a validation set at a ratio of 8:2 with the same age distribution of participants. After training the model with the development set, the model was evaluated with the test set, and this process was repeated as many times as the number of data, so that all data were used for the model evaluation. The final performance of the model was obtained by averaging each evaluation result. The regression performance of the developed model was represented as RMSE. The ANN-based regression model proposed in this study was developed using 2.90 GHz Intel Core i7-10700 processor, 64 GB 1,333 MHz DDR4 RAM, NVIDIA Geforce RTX 2070 Super, Python 3.6.7: Anaconda, and Tensorflow 2.3.0.

XSL•FO
RenderX

Table 3. Different values of hyperparameters for ANN[a]-based regression model for the estimation of vascular aging. Bold type indicates the hyperparameters for the optimal model.

Parameter	Value
Input Layer Nodes	**78**
Output Layer Nodes	**1**
Hidden Layers Number	**1** 2 3 4
Hidden Layer Nodes	64 **128** 256 512 1024
Activation Function	**ReLU[b]**
Dropout Probability	0 0.1 0.3 **0.5**
Kernel Initializer	**He_uniform**
Loss Function	**MAE[c]**
Learning Rate	0.01 0.005 **0.001** 0.0005 0.0001
Optimizer	SGD[d]**Adam**
Early Stopping Patience	30 **50**
Input Data Scaler	Standard **Robust**

[a]ANN: artificial neural network.

[b]ReLU: rectified linear unit.

[c]MAE: mean absolute error.

[d]SGD: stochastic gradient descent.

Figure 2. Architecture of the optimal version of the ANN-based regression model developed in this study. ANN: artificial neural network; INC: incident wave; OPPG: original photoplethysmogram; RPPG: reconstructed photoplethysmogram.
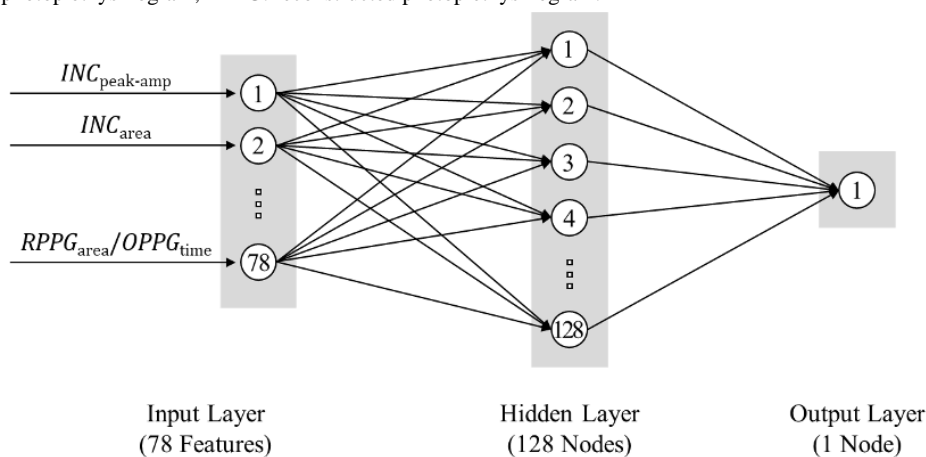
Statistical Analysis

The Pearson correlation coefficient was calculated to investigate the relationship between the participants' actual age and each feature, which was defined from the waveforms before and after the decomposition of PPG into the incident and reflected wave. The RMSE and coefficient of determination of the age estimated by the ANN-based vascular aging estimation model, which was developed with all the PPG features defined in this study, were calculated. In addition, using the estimated age and the actual age, a scatter plot and a Bland-Altman plot were made and used to analyze the model's estimation performance.

Results

Correlation Analysis

The results of the correlation analysis between the actual age and the PPG features are as follows. The correlation coefficient between the actual age and the basic features, which is defined from the original PPG, the incident and reflected waves decomposed from PPG, and the reconstructed PPG, is shown in Table 4. The reflected wave and the reconstructed PPG-related features showed a high correlation with the actual age. Among the reflected wave and reconstructed PPG features, $REF_{peak-amp}$, REF_{area}, $RPPG_{area}$, and $RPPG_{skew}$ showed a correlation greater than a weak correlation ($|R|>0.3$), and their correlation coefficients were –0.42, –0.45, and –0.45, respectively. However, most of the features defined from the

incident wave and the first inflection point showed a very weak correlation or did not have significant correlation with age. Among the features defined from the original PPG signal, the features using the diastolic peak or skewness feature of the PPG waveform showed a high correlation with age. The features showing the highest correlation in each type of pulse were $SYS_{peak-time}$, $DIA_{peak-amp}$, and $OPPG_{skew}$, and their correlation coefficients were 0.27, –0.39, and 0.41, respectively. Individual features showed a high correlation with age in the order of REF_{area}, $REF_{peak-amp}$, and $OPPG_{skew}$, and their correlation coefficient values were –0.45, –0.42, and 0.41, respectively. However, $INC_{peak-amp}$, REF_{time}, INF_{area}, $RPPG_{kurt}$, $SYS_{peak-amp}$, $OPPG_{area}$, and $OPPG_{kurt}$ showed no statistically significant correlation with the actual age, and their P values were .10, .51,

.28, .23, .52, .12, and .05, respectively. Table 5 shows the correlation between the actual age and the combined features created by the combination of the basic features. In comparing the amplitude domain feature and the temporal domain feature, some features in the spatial domain feature showed more than a weak correlation ($|R|>0.3$) with the actual age, but most of the temporal domain feature showed no correlation or only a very weak correlation ($|R|<0.3$) with the actual age. As a result, it was found that the combined feature in the spatial domain has a higher correlation with age than the combined feature in the temporal domain. Among the spatial domain features, $INC_{area}/REF_{peak-amp}$, INC_{area}/REF_{area}, and $SYS_{peak-amp}$ $-REF_{peak-amp}$ showed high correlation with age, and their correlation coefficients were 0.38, 0.37, and 0.42.

Table 4. Correlation coefficient and *P* value of basic features defined from the incident and reflected waves, first inflection point, reconstructed PPG[a], and original PPG.

Pulse type and feature	R[b]	*P* value
Incident wave		
$INC^c_{peak\text{-}amp}$	0.06	*P*=.104
INC_{area}	0.15	*P*<.001
$INC_{peak\text{-}time}$	0.23	*P*<.001
INC_{time}	0.18	*P*<.001
INC_{skew}	−0.16	*P*<.001
INC_{kurt}	−0.18	*P*<.001
Reflected wave		
$REF^d_{peak\text{-}amp}$	−0.42	*P*<.001
REF_{area}	−0.45	*P*<.001
$REF_{peak\text{-}time}$	0.10	*P*=.008
REF_{time}	0.02	*P*=.514
REF_{skew}	0.19	*P*<.001
REF_{kurt}	0.18	*P*<.001
First inflection point		
$INF^e_{peak\text{-}amp}$	−0.08	*P*=.022
$INF_{peak\text{-}time}$	0.18	*P*<.001
INF_{area}	−0.04	*P*=.275
Reconstructed PPG		
$RPPG^f_{area}$	−0.39	*P*<.001
$RPPG_{skew}$	0.40	*P*<.001
$RPPG_{kurt}$	0.04	*P*=.230
Original PPG		
$SYS^g_{peak\text{-}amp}$	0.02	*P*=.525
$SYS_{peak\text{-}time}$	0.27	*P*<.001
$DIA^h_{peak\text{-}amp}$	−0.39	*P*<.001
$DIA_{peak\text{-}time}$	0.24	*P*<.001
$OPPG^i_{area}$	0.06	*P*=.118
$OPPG_{time}$	0.08	*P*<.027
$OPPG_{skew}$	0.41	*P*<.001
$OPPG_{kurt}$	−0.07	*P*=.051

[a]PPG: photoplethysmogram.

[b]R: Pearson correlation coefficient.

[c]INC: incident wave.

[d]REF: reflected wave.

[e]INF: inflection point.

[f]RPPG: reconstructed photoplethysmogram.

[g]SYS: systolic.

[h]DIA: diastolic.

[i]OPPG: original photoplethysmogram.

Table 5. Correlation coefficient and *P* value of combined features created from the basic features.

Domain and feature	R[a]	*P* value
Spatial		
$INC^b_{peak-amp}/INC_{area}$	–0.18	*P*<.001
$INC_{peak-amp}/REF^c_{peak-amp}$	0.32	*P*<.001
$INC_{peak-amp}/REF_{area}$	0.32	*P*<.001
$INC_{peak-amp}/RPPG^d_{area}$	0.28	*P*<.001
$INC_{area}/REF_{peak-amp}$	0.38	*P*<.001
INC_{area}/REF_{area}	0.37	*P*<.001
$INC_{area}/RPPG_{area}$	0.34	*P*<.001
$REF_{peak-amp}/REF_{area}$	0.19	*P*<.001
$REF_{peak-amp}/RPPG_{area}$	–0.34	*P*<.001
$REF_{area}/RPPG_{area}$	–0.34	*P*<.001
$INC_{peak-amp}–REF_{peak-amp}$	0.31	*P*<.001
$INC_{area}–REF_{area}$	0.34	*P*<.001
$SYS^e_{peak-amp}–INC_{peak-amp}$	–0.06	*P*=.104
$SYS_{peak-amp}–REF_{peak-amp}$	0.42	*P*<.001
Temporal		
$INC_{peak-time}/INC_{time}$	0.20	*P*<.001
$INC_{peak-time}/REF_{peak-time}$	0.15	*P*<.001
$INC_{peak-time}/REF_{time}$	0.23	*P*<.001
$INC_{peak-time}/OPPG^f_{time}$	0.22	*P*<.001
$INC_{time}/REF_{peak-time}$	0.12	*P*<.001
INC_{time}/REF_{time}	0.24	*P*<.001
$INC_{time}/OPPG_{time}$	0.19	*P*<.001
$REF_{peak-time}/REF_{time}$	0.16	*P*<.001
$REF_{peak-time}/OPPG_{time}$	0.12	*P*<.001
$REF_{time}/OPPG_{time}$	–0.19	*P*<.001
$REF_{peak-time}–INC_{peak-time}$	0.03	*P*=.390
$OPPG_{time}–INC_{peak-time}$	0.03	*P*=.369
$OPPG_{time}–REF_{peak-time}$	0.03	*P*=.399
Spatiotemporal		
$INC_{peak-amp}/INC_{peak-time}$	–0.28	*P*<.001
$INC_{peak-amp}/INC_{time}$	–0.22	*P*<.001
$INC_{peak-amp}/REF_{peak-time}$	–0.11	*P*=.002
$INC_{peak-amp}/REF_{time}$	–0.02	*P*=.615
$INC_{peak-amp}/OPPG_{time}$	–0.09	*P*=.015
$INC_{area}/INC_{peak-time}$	–0.15	*P*<.001
INC_{area}/INC_{time}	–0.11	*P*=.002

Domain and feature	R[a]	P value
$INC_{area}/REF_{peak\text{-}time}$	-0.05	P=.213
INC_{area}/REF_{time}	0.02	P=.496
$INC_{area}/OPPG_{time}$	–0.02	P=.542
$REF_{peak\text{-}amp}/INC_{peak\text{-}time}$	–0.36	P<.001
$REF_{peak\text{-}amp}/INC_{time}$	–0.32	P<.001
$REF_{peak\text{-}amp}/REF_{peak\text{-}time}$	–0.22	P<.001
$REF_{peak\text{-}amp}/REF_{time}$	–0.17	P<.001
$REF_{peak\text{-}amp}/OPPG_{time}$	–0.22	P<.001
$REF_{area}/INC_{peak\text{-}time}$	–0.40	P<.001
REF_{area}/INC_{time}	–0.36	P<.001
$REF_{area}/REF_{peak\text{-}time}$	–0.27	P<.001
REF_{area}/REF_{time}	–0.25	P<.001
$REF_{area}/OPPG_{time}$	–0.28	P<.001
$RPPG_{area}/INC_{peak\text{-}time}$	–0.33	P<.001
$RPPG_{area}/INC_{time}$	–0.28	P<.001
$RPPG_{area}/REF_{peak\text{-}time}$	–0.17	P<.001
$RPPG_{area}/REF_{time}$	–0.10	P=.005
$RPPG_{area}/OPPG_{time}$	–0.16	P<.001

[a]R: Pearson's correlation coefficient.

[b]INC: incident wave.

[c]REF: reflected wave.

[d]RPPG: reconstructed photoplethysmogram.

[e]SYS: systolic.

[f]OPPG: original photoplethysmogram.

## Statistical Results of Vascular Aging Assessment

The RMSE for the age estimation of the ANN-based regression model developed in this study was 10.0 years. Figure 3 shows the scatter plot of the participant's age estimated through the ANN-based regression model corresponding to the actual age and the coefficient of determination of the model. The estimated age and actual age of the ANN-based regression model showed a high correlation of 0.63 (P<.001), and the coefficient of determination of the model was 0.4. Figure 4 shows the Bland-Altman plot for the estimated age and the actual age through the ANN-based regression model developed in this study. The upper and lower limits of 95 % agreement were 18.2 and –20.6 years, respectively.

Figure 3. Scatter plot and coefficient of determination for the ANN-based regression model developed for the estimation of vascular aging in this study. ANN: artificial neural network.
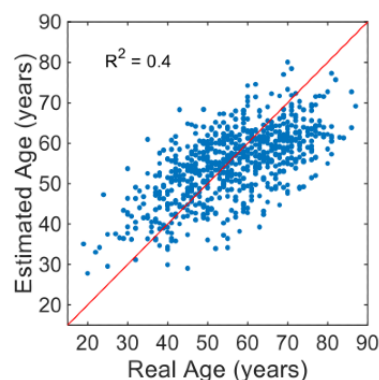
Figure 4. Bland-Altman plot for the ANN-based regression model developed for the estimation of vascular aging in this study. ANN: artificial neural network.
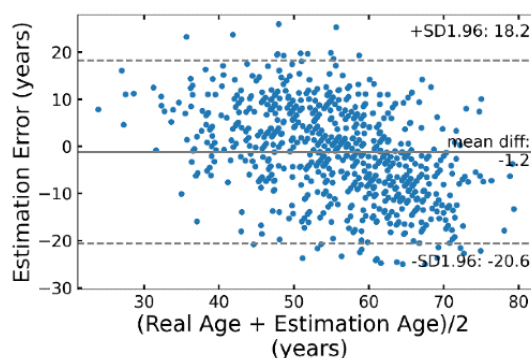
Discussion

In this study, a highly correlated feature for assessing vascular aging was explored using features before and after decomposition of the incident and reflected waves of the PPG, and an ANN-based vascular aging estimation model was developed with the features derived. The ANN-based regression model showed the RMSE of 10.0 years in the age estimation. In comparing the correlation analysis before and after decomposition of the PPG incident and reflected waves, the feature defined after decomposition rather than before decomposition of the incident and reflected waves is useful for assessing vascular aging. In addition, in the comparison of all individual features, the feature defined from the reflected wave was confirmed as the best feature for assessing vascular aging. This reconfirms that changes in arterial stiffness due to vascular aging are reflected very well in the reflected wave characteristics of PPG, as Dawber et al [13] revealed. In the comparison of the correlation between the basic features defined from the PPG original waveform, incident wave, reflected wave, and reconstructed PPG waveform before and after PPG decomposition, REF_{area} and $REF_{peak-amp}$ showed the highest correlation. These features are defined from the characteristic points of the reflected wave. In the results of our study, the time-related features of the reflected wave, such as $REF_{peak-time}$ and REF_{time}, showed a very weak correlation (P=.01) or no significant correlation (P=.51) with actual age, respectively. However, the amplitude-related feature of the reflected wave of $REF_{peak-amp}$ and REF_{area} showed a weak correlation ($|R|>0.3$) with the actual age (R=–0.42 and R=–0.45 respectively). This result suggests that the amplitude-related feature of the reflected wave is more advantageous in estimating vascular aging than the time-related feature. The result in this study is different from the study of Millasseau et al [14,15], which found that SI, an index related to the temporal characteristic of the reflected wave, had a higher correlation with age than RI, an index related to the amplitude of the reflected wave. Also, in contrast to the results of Millasseau et al and Yousef et al [14-16], the results of this study showed that the amplitude of the reflected wave decreased with aging, and the RI decreased accordingly. Unlike the previous studies that used the PPG measured from the finger, it is thought that in this study, the use of the PPG measured from the nose had an effect. In a study by Hartmann et al [22],

which observed changes in the main features of the PPG depending on the measurement location, it was reported that features such as RI could have a significant difference depending on the measurement location. Whether the change in the PPG waveform due to vascular aging has a specific pattern for each measurement location has yet to be clearly clarified; therefore, for clarification, additional research needs to be performed.

For the model development, hyperparameter optimization, such as number of hidden layers, number of nodes, dropout rate, learning rate, optimizer, early stopping patience, and input data scaler of the ANN-based regression model, was performed. In determining the hidden layer, as the number of hidden layers and the number of nodes in the hidden layer decreased, the age estimation error of the proposed model tended to decrease. In addition, as the dropout ratio of the hidden layer increased, the estimation error decreased. This means that the proposed model has sufficient expressive power to overfit the training data and that performance can be improved by suppressing overfitting [23,24]. For other training conditions used for model training, the model showed the highest performance when the optimizer was set to Adam, the learning rate was set to 0.001, and the early stopping patience was set to 50. In the case of the scaler that normalizes the input value, it was confirmed that the Robust scaler showed better performance. This is presumably because the input data contains many outliers. In comparing the performance of the model introduced in the previous study and the ANN-based regression model proposed in this study, the PPG features proposed by Millasseau et al and Yousef et al [14-16] weakly correlated with the actual age (R=–0.29 and R=–0.33 respectively). However, the ANN-based regression model proposed in this study strongly correlated with the actual age (R=0.63) and had better performance. In addition, the ANN-based regression model of this study had better performance than the previous studies in estimating the actual age of participants [18,19]. Similar to this study, the existing CNN model developed by Dall'Olio et al [18] with a single PPG input showed an estimation error of 12 years in RMSE, but the ANN-based regression model of this study showed a low estimation error of 10 years in RMSE. Moreover, our model has better estimation performance than the linear and ANN models using multiple inputs of PPG and ECG, showing estimation errors of 12 and 11 years, respectively [19] (see Table 6).

Table 6. Comparison of the proposed model to the models of previous studies in root mean squared error, correlation coefficient, and *P* value.

Reference and type of regression model	Input	RMSE[a] (years)	R	*P* value
Proposed, ANN[b]	Features from raw PPG[c] and incident and reflected wave separated from raw PPG	10	0.63	*P*<.001
Millasseau et al [14,15], linear	Feature from raw PPG	N/A[d]	–0.29	*P*<.001
Yousef et al [16], linear	Feature from raw PPG	N/A	–0.33	*P*<.001
Dall'Olio et al [18], CNN[e]	Raw PPG	12	N/A	N/A
Chiarelli et al [19]				
Linear	Feature from raw PPG and ECG[f]	12	0.64	*P*<.001
ANN	Feature from raw PPG and ECG	11	0.74	*P*<.001
DCNN[g]	Raw PPG and ECG	7	0.92	*P*<.001

[a]RMSE: root mean squared error.

[b]ANN: artificial neural network.

[c]PPG: photoplethysmogram.

[d]N/A: not applicable.

[e]CNN: convolutional neural network.

[f]ECG: electrocardiogram.

[g]DCNN: deep convolutional neural network.

This study has some limitations. Most of the previous studies that performed vascular aging evaluation used finger PPG. However, in this study, vascular aging was evaluated based on nasal PPG. Therefore, it is difficult to generalize the results of this study to a vascular aging evaluation technique using PPG regardless of the measurement location. Therefore, it is necessary to analyze the aging-related waveform change characteristics of PPG obtained from various measuring sites through additional studies. In addition, the ANN-based regression model developed in this study for estimating vascular aging is a relatively simple machine learning model with one hidden layer. Therefore, in future studies, it is necessary to improve the vascular aging estimation performance by applying a more sophisticated machine learning technique with increased model complexity. Moreover, this study did not investigate various risk factors that can accelerate vascular disease, such as atherosclerosis; therefore, it is necessary to evaluate the model performance and examine the possibility of application according to various subject characteristics.

Conclusion

In this study, we derived various features from the decomposed PPG waveforms before and after decomposition of the waveform into incident and reflected waves to explore features highly correlated with vascular aging, and it was confirmed that the reflected wave-related features had a strong correlation with participant's age. In addition, the ANN-based regression model developed using the derived feature had 10 years of RMSE in estimating the participants' actual age and showed the improved vascular aging estimation performance in comparison with the models introduced in previous studies. These results suggest that the developed technology can be applied to a wearable device and used to assess vascular health in real-life situations. However, this study was performed based on nasal PPG, not finger PPG, which is not frequently used in vascular aging evaluation studies. Since it is not clear whether the change in the PPG waveform due to vascular aging has a specific pattern for each measurement location, additional research needs to be performed for clarification.

Authors' Contributions

JP contributed to data analysis, drafting, writing, and figure design and drawing. HS contributed to the conception and design of the work, supervision, writing, and in-depth review. All authors contributed to the critical review of the final document.

Conflicts of Interest

None declared.

References

1. Laurent S, Boutouyrie P. Recent advances in arterial stiffness and wave reflection in human hypertension. Hypertension 2007 Jun;49(6):1202-1206. [doi: 10.1161/HYPERTENSIONAHA.106.076166] [Medline: 17452508]

2. Lakatta EG. Cardiovascular Regulatory Mechanisms in Advanced Age. Physiol Rev 1993 Apr;73(2):413-467. [doi: 10.1152/physrev.1993.73.2.413]

3. Lakatta EG, Levy D. Arterial and cardiac aging: major shareholders in cardiovascular disease enterprises: Part I: aging arteries: a "set up" for vascular disease. Circulation 2003 Jan 07;107(1):139-146. [doi: 10.1161/01.cir.0000048892.83521.58] [Medline: 12515756]

4. Nilsson PM, Boutouyrie P, Laurent S. Vascular aging: A tale of EVA and ADAM in cardiovascular risk assessment and prevention. Hypertension 2009 Jul;54(1):3-10. [doi: 10.1161/HYPERTENSIONAHA.109.129114] [Medline: 19487587]

5. Murgo JP, Westerhof N, Giolma JP, Altobelli SA. Aortic input impedance in normal man: relationship to pressure wave forms. Circulation 1980 Jul;62(1):105-116. [doi: 10.1161/01.cir.62.1.105] [Medline: 7379273]

6. Wilkinson IB, Fuchs SA, Jansen IM, Spratt JC, Murray GD, Cockcroft JR, et al. Reproducibility of pulse wave velocity and augmentation index measured by pulse wave analysis. J Hypertens 1998 Dec;16(12 Pt 2):2079-2084. [doi: 10.1097/00004872-199816121-00033] [Medline: 9886900]

7. McEniery CM, Yasmin, Hall IR, Qasem A, Wilkinson IB, Cockcroft JR, ACCT Investigators. Normal vascular aging: differential effects on wave reflection and aortic pulse wave velocity: the Anglo-Cardiff Collaborative Trial (ACCT). J Am Coll Cardiol 2005 Nov 01;46(9):1753-1760 [FREE Full text] [doi: 10.1016/j.jacc.2005.07.037] [Medline: 16256881]

8. Nilsson PM, Khalili P, Franklin SS. Blood pressure and pulse wave velocity as metrics for evaluating pathologic ageing of the cardiovascular system. Blood Press 2014 Feb;23(1):17-30. [doi: 10.3109/08037051.2013.796142] [Medline: 23750722]

9. Antza C, Doundoulakis I, Akrivos E, Stabouli S, Trakatelli C, Doumas M, et al. Early vascular aging risk assessment from ambulatory blood pressure monitoring: The early vascular aging ambulatory score. Am J Hypertens 2018 Oct 15;31(11):1197-1204. [doi: 10.1093/ajh/hpy115] [Medline: 30239585]

10. Rolfe P. Photoelectric plethysmography for estimating cutaneous blood flow. New York: Academic; 1979:125-151.

11. Millasseau SC, Guigui FG, Kelly RP, Prasad K, Cockcroft JR, Ritter JM, et al. Noninvasive assessment of the digital volume pulse. Comparison with the peripheral pressure pulse. Hypertension 2000 Dec;36(6):952-956. [doi: 10.1161/01.hyp.36.6.952] [Medline: 11116106]

12. Millasseau SC, Ritter JM, Takazawa K, Chowienczyk PJ. Contour analysis of the photoplethysmographic pulse measured at the finger. J Hypertens 2006 Aug;24(8):1449-1456. [doi: 10.1097/01.hjh.0000239277.05068.87] [Medline: 16877944]

13. Dawber TR, Thomas HE, McNamara PM. Characteristics of the dicrotic notch of the arterial pulse wave in coronary heart disease. Angiology 1973 Apr;24(4):244-255. [doi: 10.1177/000331977302400407] [Medline: 4699520]

14. Millasseau SC, Kelly RP, Ritter JM, Chowienczyk PJ. Determination of age-related increases in large artery stiffness by digital pulse contour analysis. Clin Sci (Lond) 2002 Oct;103(4):371-377. [doi: 10.1042/cs1030371] [Medline: 12241535]

15. Millasseau SC, Kelly RP, Ritter JM, Chowienczyk PJ. The vascular impact of aging and vasoactive drugs: comparison of two digital volume pulse measurements. Am J Hypertens 2003 Jun;16(6):467-472. [doi: 10.1016/s0895-7061(03)00569-7] [Medline: 12799095]

16. Yousef Q, Reaz MBI, Ali MAM. The analysis of PPG morphology: Investigating the effects of aging on arterial compliance. Meas Sci 2012 Jan 1;12(6):266-271. [doi: 10.2478/v10048-012-0036-3]

17. Park J, Shin H. Development of vascular aging assessment model based on photoplethysmogram incident and reflected wave characteristics. Trans Korean Inst Electr Eng 2021 Apr 30;70(4):700-706. [doi: 10.5370/kiee.2021.70.4.700]

18. Dall'Olio L, Curti N, Remondini D, Safi Harb Y, Asselbergs FW, Castellani G, et al. Prediction of vascular aging based on smartphone acquired PPG signals. Sci Rep 2020 Nov 12;10(1):19756 [FREE Full text] [doi: 10.1038/s41598-020-76816-6] [Medline: 33184391]

19. Chiarelli AM, Bianco F, Perpetuini D, Bucciarelli V, Filippini C, Cardone D, et al. Data-driven assessment of cardiovascular ageing through multisite photoplethysmography and electrocardiography. Med Eng Phys 2019 Nov;73:39-50. [doi: 10.1016/j.medengphy.2019.07.009] [Medline: 31358395]

20. Reynolds DA. Speaker identification and verification using Gaussian mixture speaker models. Speech Commun 1995 Aug;17(1-2):91-108. [doi: 10.1016/0167-6393(95)00009-d]

21. MASS AL, Hannum AY, Ng AY. Rectifier nonlinearities improve neural network acoustic models. 2013 Presented at: International conference on machine learning; 16-21 June 2013; Atlanta, Georgia, USA. [doi: 10.1097/gme.0b013e3181846cb6]

22. Hartmann V, Liu H, Chen F, Qiu Q, Hughes S, Zheng D. Quantitative Comparison of Photoplethysmographic Waveform Characteristics: Effect of Measurement Site. Front Physiol 2019;10:198 [FREE Full text] [doi: 10.3389/fphys.2019.00198] [Medline: 30890959]

23. Huang GB, Zhu QY, Siew CK. Extreme learning machine: Theory and applications. Neurocomputing 2006 Dec;70(1-3):489-501. [doi: 10.1016/j.neucom.2005.12.126]

24. Fletcher L, Katkovnik V, Steffens FE, Engelbrecht AP. Optimizing the number of hidden nodes of a feedforward artificial neural network. 1998 Presented at: 1998 IEEE International Joint Conference on Neural Networks Proceedings IEEE World

Congress on Computational Intelligence (Cat No 98CH36227); 4-9 May 1998; Anchorage, Alaska, USA. [doi: 10.1109/ijcnn.1998.686018]

Abbreviations

AIx: augmentation index
ANN: artificial neural network
CNN: convolutional neural network
DCNN: deep convolutional neural network
DIA: diastolic
ECG: electrocardiogram
INC: incident wave
INF: inflection point
LOOCV: leave-one-out cross-validation
MAE: mean absolute error
OPPG: original photoplethysmogram
PPG: photoplethysmogram
REF: reflected wave
ReLU: rectified linear unit
RI: reflection index
RMSE: root mean squared error
RPPG: reconstructed photoplethysmogram
SI: stiffness index
SYS: systolic

Original Paper

Prediction of Chronic Obstructive Pulmonary Disease Exacerbation Events by Using Patient Self-reported Data in a Digital Health App: Statistical Evaluation and Machine Learning Approach

Francis P Chmiel[1], MSc, DPhil; Dan K Burns[1], MSc, PhD; John Brian Pickering[1], DPhil; Alison Blythin[2], MRES; Thomas MA Wilkinson[2,3,4*], PhD; Michael J Boniface[1*], BEng

[1]School of Electronics and Computer Science, University of Southampton, Southampton, United Kingdom

[2]my mHealth Limited, Bournemouth, United Kingdom

[3]National Institute for Health Research Applied Research Collaboration Wessex, University of Southampton, Southampton, United Kingdom

[4]Faculty of Medicine, University of Southampton, Southampton, United Kingdom

*these authors contributed equally

Corresponding Author:
Francis P Chmiel, MSc, DPhil
School of Electronics and Computer Science
University of Southampton
University Road
Southampton, SO17 1BJ
United Kingdom
Phone: 44 023 8059 8866
Email: F.P.Chmiel@soton.ac.uk

Abstract

Background: Self-reporting digital apps provide a way of remotely monitoring and managing patients with chronic conditions in the community. Leveraging the data collected by these apps in prognostic models could provide increased personalization of care and reduce the burden of care for people who live with chronic conditions. This study evaluated the predictive ability of prognostic models for the prediction of acute exacerbation events in people with chronic obstructive pulmonary disease by using data self-reported to a digital health app.

Objective: The aim of this study was to evaluate if data self-reported to a digital health app can be used to predict acute exacerbation events in the near future.

Methods: This is a retrospective study evaluating the use of symptom and chronic obstructive pulmonary disease assessment test data self-reported to a digital health app (myCOPD) in predicting acute exacerbation events. We include data from 2374 patients who made 68,139 self-reports. We evaluated the degree to which the different variables self-reported to the app are predictive of exacerbation events and developed both heuristic and machine learning models to predict whether the patient will report an exacerbation event within 3 days of self-reporting to the app. The model's predictive ability was evaluated based on self-reports from an independent set of patients.

Results: Users self-reported symptoms, and standard chronic obstructive pulmonary disease assessment tests displayed correlation with future exacerbation events. Both a baseline model (area under the receiver operating characteristic curve [AUROC] 0.655, 95% CI 0.689-0.676) and a machine learning model (AUROC 0.727, 95% CI 0.720-0.735) showed moderate ability in predicting exacerbation events, occurring within 3 days of a given self-report. Although the baseline model obtained a fixed sensitivity and specificity of 0.551 (95% CI 0.508-0.596) and 0.759 (95% CI 0.752-0.767) respectively, the sensitivity and specificity of the machine learning model can be tuned by dichotomizing the continuous predictions it provides with different thresholds.

Conclusions: Data self-reported to health care apps designed to remotely monitor patients with chronic obstructive pulmonary disease can be used to predict acute exacerbation events with moderate performance. This could increase personalization of care by allowing preemptive action to be taken to mitigate the risk of future exacerbation events.

XSL·FO
RenderX

Introduction

Chronic obstructive pulmonary disease (COPD) is a collection of progressive lung diseases, characterized by breathing difficulties and an irreversible reduction of lung function. It is one of the most prevalent chronic conditions in the world (in England, 2.19% of the population is expected to have a confirmed COPD diagnosis by 2030 [1]), and the absence of a cure means it represents a significant burden for patients who have to manage the condition on a daily basis [2,3]. A key characteristic of managing COPD is in mitigating the risk of "exacerbation events," which can be defined as an acute sustained worsening of a patient's condition that necessitates a change in medication or emergency care, including hospitalization [4]. Exacerbations accelerate lung function decline, and evidence suggests that the frequency of exacerbations increases with decreasing lung function [5-7]. Minimizing the number of exacerbation events can therefore have a significant impact on the prognosis for patients with COPD. Currently, several methods exist to help control exacerbation events, including pharmacological interventions, pulmonary rehabilitation, and self-management programs [8]. There is also an identified clinical need to predict exacerbation events in advance to personalize COPD treatment and offer the opportunity to provide targeted preemptive interventions [9,10].

In recent years, the advent of mobile health apps has facilitated increased remote management and care of patients with COPD [11,12]. These apps support the recording of temporally dense information about a patient's condition, which allow (near) real-time monitoring of a patient's symptoms, providing clinicians with a source of data to help them understand how the patient is managing their condition and gain an insight into the patient's exacerbation frequency and severity. For the patient, digital health apps provide both an access point for educational content about their condition and the opportunity to improve their self-care, leading to better long-term management [13]. In the context of COPD, there is an opportunity to increase the efficacy of digital health apps further by leveraging the data they collect to predict acute exacerbation events and provide personalized alerts to the patient. These alerts could facilitate a clinically validated and personalized intervention program to mitigate the occurrence and reduce the severity of an acute exacerbation event.

In this report, we present a retrospective study making use of data collected by the myCOPD mobile app, a National Health Service–approved, clinically validated app for persons with a diagnosis of COPD [14]. This app assists with the management of COPD by providing educational content alongside a digital momentarily assessed symptom diary. Using this app, users self-report on the COPD-related symptoms they are currently experiencing as well as information that characterizes their long-term COPD status (ie, the COPD Assessment Test [CAT]). Using statistical analysis and machine learning methods, we evaluate the effectiveness of exploiting this simple self-reported information to predict exacerbation events in the near future and discuss how such predictions could be used to improve the long-term outcomes for people with COPD.

Methods

Data Set Description

We used an anonymized extract of daily user self-reports submitted to the myCOPD app between January 1, 2017 and December 31, 2019 (inclusive). All users of the myCOPD app are clinically diagnosed with COPD, with app usage limited to patients "prescribed" the app by clinicians as part of agreed care plans. A single report features a self-assessed symptom score (Figure 1), which is a 4-point scale ranging from normal symptoms to a severe deterioration of symptoms requiring medical intervention, including hospitalization. We encode symptoms as an ordinal variable between 1 and 4, indicating increasing severity of COPD-related symptoms (Figure 1A). Symptom scores have a level of subjectivity across users (eg, the severity of symptoms considered normal for a given user will vary) with users provided with education to increase awareness and understanding of what baseline symptom scores would be considered normal for them. Users also perform a CAT at regular (approximately monthly) intervals. The CAT is an 8-question assessment and yields a score between 0 and 40, where higher values indicate a more severe impact of COPD on a user's overall health [15]. CAT is a validated and an accepted way of quantifying the burden of COPD on someone's life [16,17]. The reporting frequency of symptoms and CAT scores for our cohort is presented in Table S1 in Multimedia Appendix 1. In addition to these scores, our data set also features additional demographic and lifestyle information self-reported to the app. These include patient age, gender, current smoking status, and the number of years they have been smoking for.

Figure 1. Self-reported symptom scores and chronic obstructive pulmonary disease assessment test (CAT) scores. (A) Symptom score rankings and classification of whether this score corresponds to an exacerbation event, as defined in the context of this work. (B) Example user (with high reporting frequency) self-reporting timeline where the top panel displays self-reported symptom scores and the bottom panel self-reported CAT results. CAT: chronic obstructive pulmonary disease assessment test.
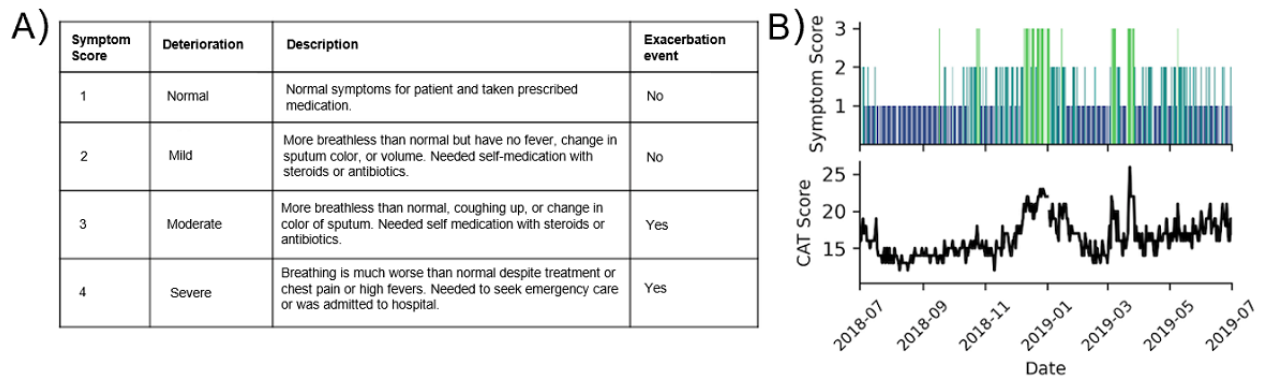
Ethics Approval and Data Governance

This work received ethics approval from the University of Southampton's Faculty of Engineering and Physical Science Research Ethics Committee (ERGO/FEPS/52137) and was reviewed by the University of Southampton Data Protection Impact Assessment panel (DPIA 0045), with the decision to support the research.

Defining Exacerbation Events

We use a symptom-based definition for exacerbation events where an event is marked to have occurred if a patient self-reports a score of 3 or 4 corresponding to a moderate or severe deterioration, respectively, from a patient's normal symptoms. A score of 3 indicates that a patient is more breathless than normal, coughing up sputum or with change in sputum color, and has needed to self-medicate using steroids or antibiotics. A score of 4 indicates that a patient's breathing is much worse than normal despite treatment, has chest pain or a high fever, and has needed to seek emergency care or was admitted to the hospital (Figure 1) [13].

Cohort Selection and Data Set Segregation

In total, 5170 users were included in the extract who reported a total number of 94,882 reports in the study period. User registration was incremental (ie, not all users registered at the same time) throughout the period of the study, and self-reports were not necessarily submitted every day. To create our study cohort, we followed a selection process outlined in Figure 2.

First, isolated symptom reports (those in which a second report was not made within 3 days) were removed because the target variable could not be reliably calculated. Next, reports from anomalous users (those only reporting exacerbation events or entering self-reports before their registration date) were removed. After removal of these reports, we obtained our final study cohort featuring 68,139 self-reported symptom scores from 2374 unique users. Patient information relates to the time at which the patient first reported to the app (eg, if there smoking status changed, Table 1 summarizes the first reported status). Table 1 presents the characteristics of 2374 unique patients in our cohort (including patients in both train and test). For our user cohort, the mean reporting frequency between patients' first and last reports to the app was 3.28 symptom score reports per week and 0.68 CAT score reports per week.

From our cohort of 2374 users, 1672 users were between the ages of 60 years and 79 years inclusive (Table 1). Only 650 users reported their gender, with 419 males reporting compared to 231 females. A large fraction (n=1157) of users reported their smoking history, with 86.5% (1001/1157) of those reporting being either a current or ex-smoker (Table 1). Out of the self-reports included in this study, 742 patients reported 5906 self-reports that correspond to an exacerbation event, corresponding to 8.7% (5906/68,139) of the total reports and 31.3% (742/2374) of the patients (Figure 2). The median number of exacerbation event reports per patient was 3 (IQR 1-7) for our cohort.

Figure 2. Selection of self-reports in our study cohort containing 2374 patients. Isolated reports (n=24,801) were those without a subsequent report in the following 3 days. Anomalous users (n=1942) were those who only reported exacerbation events or self-reported to the myCOPD app before their registration date. Exacerbation events (n=5906) were all self-reported to the app by 742 patients.
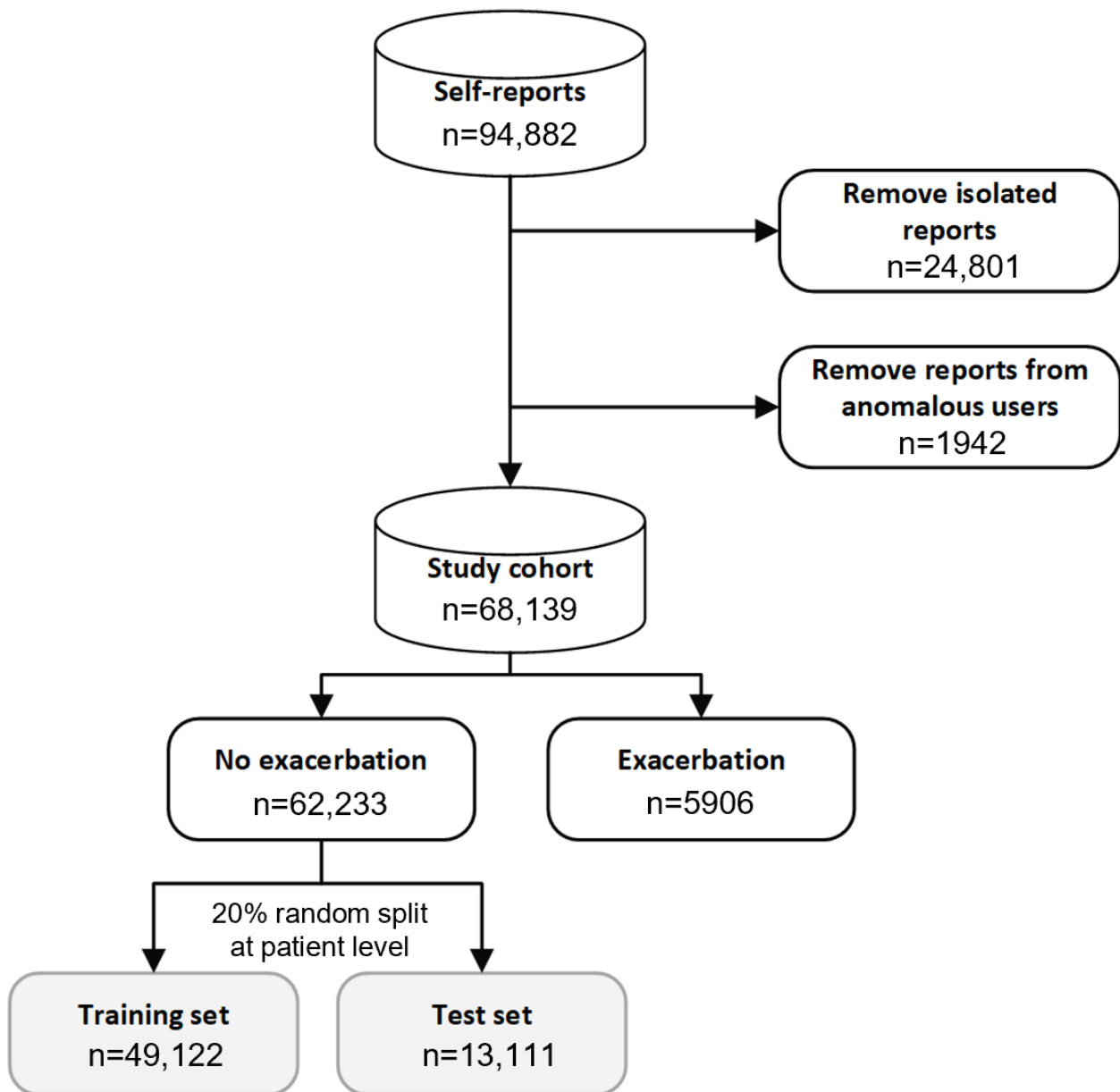
Table 1. Patient demographics and smoking status in our cohort (N=2374). All information was self-reported to the myCOPD app.

Group, subgroups	Patients, n (%)
Age group (years)	
Missing	10 (0.4)
19-29	7 (0.3)
30-39	39 (1.6)
40-49	89 (3.7)
50-59	325 (13.7)
60-69	791 (33.3)
70-79	881 (37.1)
80-89	212 (8.9)
90-99	15 (0.6)
100-110	5 (0.2)
Gender	
Missing	1724 (72.6)
Male	419 (17.6)
Female	231 (9.7)
Smoking status	
Missing	1217 (51.3)
Ex-smoker	843 (35.3)
Nonsmoker	156 (6.6)
Smoker	158 (6.7)

Predicting Exacerbation Events

For each daily self-report, we created a binary variable that indicated whether a report is followed by an exacerbation event in the following 3 days. A 3-day window was chosen empirically based on clinical guidance to be close enough to the future exacerbation event for any signal to be present in the data but sufficiently far from the event such that a range of preemptive actions could be available to patients. For the training of the prognostic models, we selected only reports in which the patient did not report an exacerbation event on the same day (n=49,122, Figure 2). We then randomly assigned reports from 19.2% (13,111/68,139) of the patients to a holdout test set (Figure 2).

We created a baseline heuristic model that uses only a user's most recently reported symptom score. The model assigns users to 2 risk groups: users reporting a symptom score of 1 are predicted to be at low risk of exacerbation (1.7% risk) within 3 days, and users reporting a symptom score of 2 are predicted to be at heightened risk (7.2% risk) of exacerbation within 3 days. Percentages in brackets correspond to the mean 3-day exacerbation rate for all reports in the training set with symptom scores of 1 or 2, respectively. The heuristic model is equivalent to a decision tree with a depth of 1. Supervised machine learning models make use of patient demographics, lifestyle information, self-reported information, and aggregate features that summarize a patient's (recent) self-reporting history. A full schema of variables used by our models is presented in Table S2 of Multimedia Appendix 1. We used logistic regression with

regularization and a random forest classifier each trained by 5-fold and grouped cross-validation at the user level, that is, reports from a single user appear exclusively in either the training or validation fold. Missing CAT scores were forward-filled imputed at the user level where possible. All other missing values were filled using mean imputation within fold. Either target or ordinal encoding was used for all categorical variables (Table S2 in Multimedia Appendix 1). Model hyperparameters were optimized on the out-of-fold validation samples by Bayesian optimization via the Tree Parzen Estimator algorithm as implemented in the HyperOpt Python library [18,19]. Model performance was evaluated on the holdout test set, and 95% CIs were estimated by bootstrapping. To create a binary decision of exacerbation risk, model predictions were dichotomized with thresholds chosen to yield either a fixed specificity or the maximum Youden's J statistic on the test set [20].

Results

### Relationship Between Self-reported Scores and Exacerbation Events

Figure 3 investigates the relationship between symptom scores and CAT scores self-reported to the myCOPD app and self-reported exacerbation events. Panels A to D of Figure 3 display the correspondence between symptom scores and the CAT results when self-reported on the same day. Although for each symptom score, users nearly report the full range of CAT

scores, there is a clear correlation between CAT scores and symptom scores, with users reporting higher symptom scores more likely to also report a higher CAT score. For example, the mean CAT score reported when a user reports a symptom score of 1 is 13.5, which is significantly lower (*P*<.001) than the mean CAT score (19.5) when a user reports a symptom score of 2. Such a correlation is to be expected; research has shown increased CAT scores correlate with an increased exacerbation frequency [21], which in turn would lead to higher symptom scores being reported in our study.

In Figures 3E and F, we evaluate the sensitivity of symptom scores and CAT results to future exacerbation events. We display how the mean of the 2 reported variables change in days preceding (and following) the day in which users self-report their first exacerbation event to the app (not necessarily their first ever exacerbation event). By inspection of Figure 3E, we see that the mean reported symptom score in proximity of the

first reported exacerbation event increases, indicating that in the days preceding their first exacerbation event, users are increasingly likely to report a mild deterioration of symptoms (symptom score of 2). Changes in the mean symptom score calculated across all users can be seen several days in advance, suggesting that at least some users are observing a mild deterioration in symptoms several days in advance of exacerbation events. For days subsequent to users' first self-reported exacerbation (right of the dashed line in Figure 3E), the mean symptom score initially exceeds 2, showing that self-reported exacerbation events can be multiday events—consistent with the current understanding of exacerbation events [4]. Similarly, in Figure 3F, the reported mean CAT result is observed to increase in magnitude in the days preceding an exacerbation event and then decrease (at a slower rate) following a reported event. Overall, these trends indicate there is potential in using these self-reported variables to predict at least a subset of exacerbation events in advance.

Figure 3. Self-reported symptom scores and results of chronic obstructive pulmonary disease assessment test (CAT) for reports in our 2374 patient cohort. (A-D) Displays the self-reported CAT result stratified by the self-reported symptom score (row) on the day of test completion. (E) Mean self-reported symptom scores in the days preceding (and following) a day where a patient self-reports their first exacerbation event. (F) Mean self-reported result of CAT in the days preceding (and following) a day where a patient self-reports their first exacerbation event. Grey dashed lines in all panels highlight the day of the first reported exacerbation event (time=0 days). Panels E and F indicate that exacerbation events can be associated with a worsening of symptom scores and CAT results several days in advance of the event. The width of the observed peaks (see panel E, right of dashed line) following the start of the exacerbation event demonstrates that exacerbation events can be multiple day events. CAT: chronic obstructive pulmonary disease assessment test.
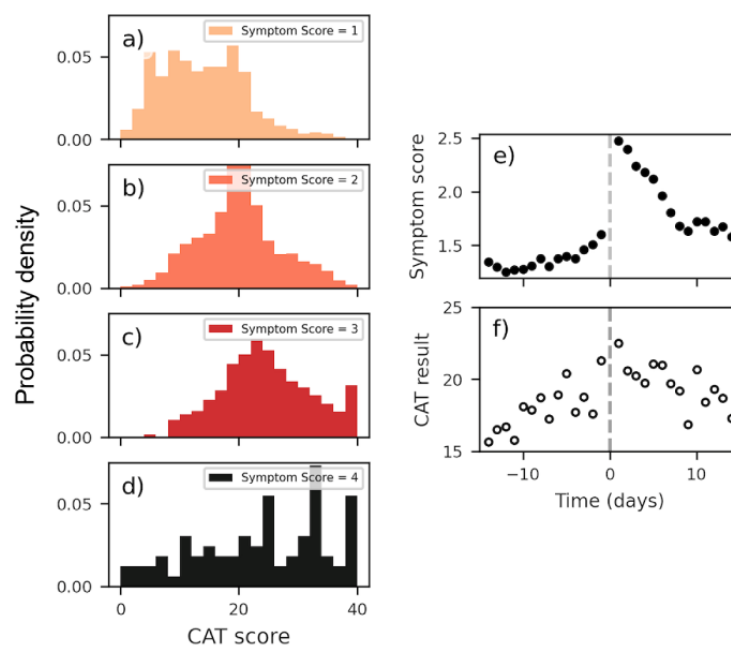

Predicting Exacerbation Events

Figure 4 shows the receiver operating characteristic (ROC) curve comparing a set of prognostic models to predict exacerbation. This includes a baseline model alongside the 2 machine learning models—a logistic regression model (solid grey line) that does not consider variable interactions and a random forest classifier (solid black line) that does. The baseline prognostic model captures the key feature about exacerbation events observed in our data: persons reporting a mild deterioration of symptoms are significantly (*P*<.001) more likely to experience an exacerbation event in the next 3 days compared

to those reporting normal symptoms, with a relative risk of 4.16 (95% CI 3.8-4.5). On the holdout test set, the baseline model obtained an area under the receiver operating characteristic (AUROC) of 0.655 (95% CI 0.676-0.689). The logistic regression model obtained an AUROC of 0.697 (95% CI 0.689-0.711) and the random forest model 0.727 (95% CI 0.720-0.735) on the holdout test (Table 2). The significantly higher (*P*<.001) performance of the random forest model suggests either interactions between variables are important in discriminating between reports associated with exacerbation within 3 days or nonlinear relations are present.

Figure 4. Model performance evaluated on the patient holdout test set. (A) Receiver operating characteristic curve of our models. The baseline model (dashed line) has only 1 nontrivial threshold for dichotomizing the prediction (diamond marker), whereas the machine learnt models has a number of possible thresholds, which needs to be optimized to suit the use case (so called sensitivity-specificity trade-off). (B) Feature importance (Gini importance) for the random forest model. CAT: chronic obstructive pulmonary disease assessment test.
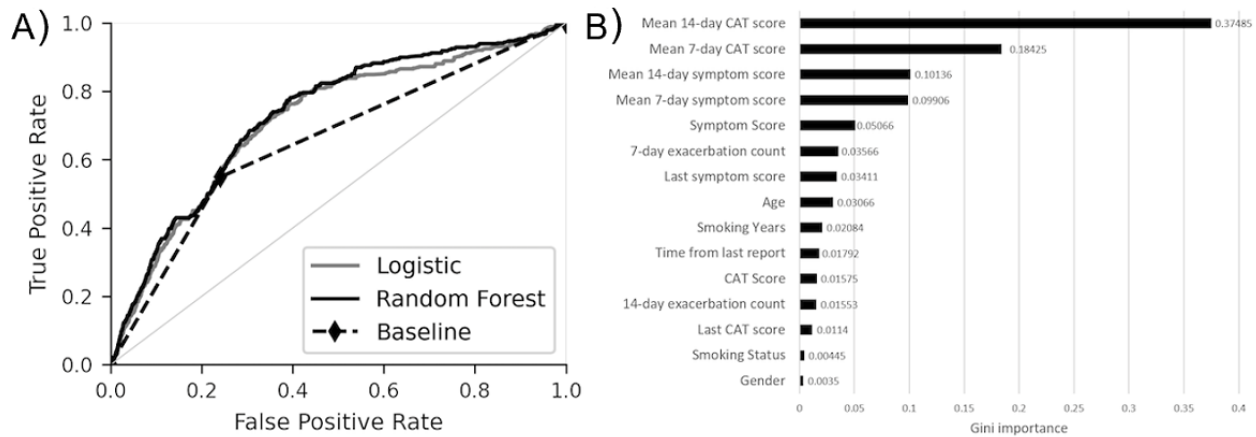
In Figure 4B, we present the feature importance (Gini importance) for our random forest model. The most important features are the patients' recent CAT scores (mean 14-day CAT score and mean 7-day CAT score), consistent with research that showed CAT scores are an effective way of quantifying the severity of a patient's COPD, which in turn, is linked to their exacerbation risk [16,17]. The next most important features are those quantifying patients' recently reported symptom scores. Symptom scores reflect the symptoms a patient is (or was recently) experiencing, and we have shown (Figure 3E) that people reporting higher symptom scores are more likely to report an exacerbation event within 3 days compared to those reporting lower symptom scores. It, therefore, is reasonable that the machine learning model can use this information to better quantify a patient's exacerbation risk.

In Table 2, we present the sensitivity and specificity of the baseline model and the machine learning models evaluated on the holdout test set. Although the baseline model is already dichotomized, for the machine learning models, a threshold must be chosen to binarize the continuous exacerbation risks they produce. The baseline model obtained a sensitivity of 0.551 (95% CI 0.508-0.596) with specificity of 0.759 (95% CI 0.752-0.767). Although neither machine learning model significantly outperforms the baseline model at the same specificity (eg, compare models A and E in Table 2), the tuning of the threshold used to dichotomize the machine learning model predictions can lead to a range of sensitivities and specificities (compare models C, D, and E in Table 2) on the holdout test, which could be tuned to match different escalation policies and interventional strategies. For example, the random forest model can be tuned to yield a sensitivity of 0.921 (95% CI 0.907-0.935) or 0.576 (95% CI 0.553-0.594) with respective specificities of 0.250 (95% CI 0.246-0.254) or 0.750 (95% CI 0.749-0.751).

Table 2. Model performances evaluated on the holdout test set.[a]

Name	Model	Area under the receiver operating characteristic curve (95% CI)	Threshold	Sensitivity (95% CI)	Specificity (95% CI)
A	Baseline model	0.655 (0.632-0.676)	N/A[b]	0.551 (0.508-0.596)	0.759 (0.752-0.767)
B	Logistic regression	0.697 (0.689-0.711)	Youden's J statistic	0.708 (0.625-0.768)	0.644 (0.574-0.706)
C	Random forest	0.727 (0.720-0.735)	Youden's J statistic	0.755 (0.676-0.813)	0.629 (0.564-0.700)
D	Random forest	0.727 (0.720-0.735)	Specificity=0.25	0.921 (0.907-0.935)	0.250 (0.246-0.254)
E	Random forest	0.727 (0.720-0.735)	Specificity=0.75	0.576 (0.553-0.594)	0.750 (0.749-0.751)

[a]The area under the receiver operating characteristic curve column denotes the area under the receiving operator curve (Figure 4) for each model. The 3 rightmost columns display the sensitivity and specificity of models at predicting exacerbations with different thresholds used to dichotomize the predictions. The baseline model is already binary and only has 1 nontrivial configuration, but the threshold used to dichotomize the machine learning models (B-E) can be tuned to suit the intended context of the model. The maximum of Youden's J statistic is used as a baseline criterion for dichotomizing the prediction (models B and C), and other cutoffs yielding fixed specificities are investigated for the random forest model. The area under the receiver operating characteristic curve for models C, D, and E are the same since they correspond to the same underlying model.

[b]N/A: not applicable.

Discussion

The etiology of COPD exacerbations is now well understood, with bacterial and viral infections, exposure to extreme weather, and air pollution being the key drivers set against the background of poor disease control. Effective treatments are available for COPD, with early recognition of symptoms of deterioration and prompt intervention having been shown to be

associated with better outcomes [22]. Current models of care for people living with COPD include the use of regular inhaled medications and rescue packs of antibiotics and steroids that they keep at home on standby for use during exacerbations. Patients are expected to initiate treatments if they suspect they have an exacerbation. To influence clinical outcome beneficially, a predictive model needs to enable a preemptive intervention with the likely impact maximized the earlier this occurs. In the context of COPD, exacerbation interventions may include increased use of inhaled medication or additional administration of rescue packs of oral antibiotics and corticosteroids [23]. The current standard of care has been for patients to take these treatments when the symptoms are exacerbating, which creates a reactive model of care requiring significant clinical deterioration to have occurred before an intervention is started. Our own work has shown that early treatment of exacerbations is associated with improved clinical outcomes, including faster recovery times [24]. As such, there is, however, a strong evidence base to suggest that this paradigm of care is inadequate and leads to increasing numbers of hospital admissions, unscheduled visits to primary care, and prolonged episodes of ill health and sickness absence from work.

Early prediction of COPD exacerbation events has the potential to change clinical practice and transform management of COPD. With the preemptive warning of a future exacerbation, the new app codeveloped with patients with COPD will alert patients of the risk and provide information on appropriate therapy options. More effective treatment of exacerbations offers the possibility of faster recovery, less relapses, and overall, therefore, better health-related quality of life, improved disease control, and potentially fewer exacerbations. Our study has shown that symptom information self-reported to the myCOPD app displayed correlation to the start of the future exacerbation events (Figure 2E), and we found that machine learning models utilizing this information and other sources of self-reported data were able to identify patients at risk of exacerbation within 3 days with moderate discriminative ability (AUROC 0.727, 95% CI 0.720-0.735). Therefore, if presented appropriately, this risk prediction model could enable patients to self-manage more effectively by intervening before life-threatening inflammation and infection can become established.

Various approaches to continuous remote monitoring of patients with COPD in communities have emerged in recent years that use sensor technologies to measure physiological parameters (respiratory rate, pulse oximetry, spirometry, blood pressure, weight, etc) and physical activity [25]. The consequence is that many of the symptoms and parameters of COPD exacerbation [26] previously measured in clinical settings can now be measured at home. This trend is transforming decision support from tools used by clinicians to tools empowering patients in everyday life. Traditionally, such tools are developed using predefined rules applied to population-based thresholds on parameters, as per our baseline model. However, new approaches are needed to support a care paradigm shifting to remote monitoring of complex parameters with varying degrees of reporting compliance, data quality and patient condition, and behaviors.

Machine learning models are well positioned to address these requirements because they can dynamically learn complex nonlinear relations between variables, which are inaccessible to handcrafted models, and despite the complexity of the models, they can be easily integrated into digital health apps such as myCOPD. This greatly facilitates the potential uptake of the model since they can be directly implemented in an active digital ecosystem for patient care that fosters data acquisition and can position predictions within new models of patient-to-clinician interaction with predictions updated every time a user provides new data. For machine learning models, it is important to match their configuration to the escalation policy. If models are used as a binary alert system, this is achieved by analyzing the (so-called) sensitivity-specificity trade-off, considered in Table 2 or Figure 4. In Table 2, 3 configurations of the random forest model are chosen (models C, D, and E), which yield different sensitivities and specificities. For example, model D in Table 2 uses a threshold chosen to obtain a specificity of 0.25 on the test set and achieves a sensitivity of 0.921 (95% CI 0.907-0.935). This configuration could be appropriate if false positives are not of significant concern (eg, if the prescribed intervention plan is of little risk to the patient). Ultimately, different configurations allow flexibility in the resulting escalation policies and is the key advantage of the machine learning models compared to the baseline model. Our use of decision tree–based algorithms offers high explainability necessary for interpretation and transparency in predictions. Single decision tree classifiers are directly explainable and provide decision support in a manner similar to classical clinical decision support tools, while ensemble methods can be made explainable by taking advantage of recent advancements (eg, Shapley additive explanations [27]).

The predictive performance of the machine learning models for COPD exacerbation can be improved by adding COPD-related variables. Health and lifestyle activity factors (including comorbidity and socioeconomic status) are believed to impact COPD exacerbation frequency [22]. These could be acquired from a patient's medical record or collected with questionnaires and used by the algorithm to further refine its predictions. Additionally, since self-reported deterioration in symptoms can occur several days before an exacerbation (Figure 2E), it is reasonable that symptom information could be collected in a more granular manner in the days preceding an exacerbation event. This could be achieved with medical devices (eg, smart inhalers) that automatically incorporate spirometry, wearables designed to monitor a person's lung function, and COPD-related physiological and behavioral variables (eg, oxygen saturation, respiration rate, temperature) or through more granular self-reporting of symptoms through the digital app [28].

Reporting compliance is a concern for safety, effectiveness, and acceptance of models. There is a need to ensure the burden of technology is reduced to a minimum, that incentives and benefits of reporting are aligned, and where appropriate, automatic observations and measurements are used to capture data (eg, smart inhalers). Our self-reports were collected in a prospective fashion by using momentary assessments of a patient's COPD symptoms. Reporting compliance and individual variation in reporting behaviors could still be detrimental to our results, with exacerbation frequencies or severity being misreported [29].

Our decision to remove isolated reports outside of the 3-day window may introduce bias owing to variations in patient reporting behaviors and the possibility of underrepresentation of exacerbating symptoms in patients who may be too ill to report at sufficient frequency, especially those admitted to the hospital. Integration with medical records would address this concern by providing the ground truth for severe exacerbation where patients require urgent medical care. Imperfect reporting compliance also acts to limit the amount of data available to our machine learning model, which may limit their predictive ability. Although accountability to clinicians may improve compliance for some patients, ideally what is required is trust, acceptance, and engagement by those people living with COPD.

Therefore, our model must be integrated into the digital health platform with patient groups participating in the co-design and optimization of the intervention to identify barriers to intervention and target behaviors prior to and during a clinical trial [30].

To conclude, our results suggest that data self-reported to a digital health app, designed for the management of people with COPD, can be used to identify users at risk of exacerbation within 3 days with moderate discriminative ability (AUROC 0.727, 95% CI 0.720-0.735). Further research utilizing additional linked data (particularly from medical devices such as smart inhalers, physiological monitoring sensors, and environmental sensors) are expected to increase the accuracy of these models.

Acknowledgments

Data Availability

Data will be made available upon reasonable request to persons with a university affiliation. Requestors will need appropriate data protection, governance, and ethical review in place.

Authors' Contributions

FPC performed the analysis with support from DKB and MJB. FPC wrote the first draft of the manuscript. FPC, JBP, and MJB wrote the second draft of the manuscript. MJB and FPC wrote the final draft of the manuscript. JBP obtained ethical and governance approvals. FPC, JBP, and MJB led the research project at University of Southampton. AB managed the data extraction at my mHealth. TMAW and AB provided clinical insight. FPC, MJB, and TMAW envisaged the research. All other authors contributed to future iterations of the manuscript.

Conflicts of Interest

TMAW is Chief Science Officer and cofounder of my mHealth, the developer of the myCOPD app. AB is a Senior Research Nurse and Clinical Trial Manager at my mHealth. All other authors declare no competing interests.

Multimedia Appendix 1
Supplementary data.
[DOCX File , 17 KB - medinform_v10i3e26499_app1.docx]

References

1. McLean S, Hoogendoorn M, Hoogenveen RT, Feenstra TL, Wild S, Simpson CR, et al. Projecting the COPD population and costs in England and Scotland: 2011 to 2030. Sci Rep 2016 Sep 01;6:31893 [FREE Full text] [doi: 10.1038/srep31893] [Medline: 27583987]
2. Lozano R, Naghavi M, Foreman K, Lim S, Shibuya K, Aboyans V, et al. Global and regional mortality from 235 causes of death for 20 age groups in 1990 and 2010: a systematic analysis for the Global Burden of Disease Study 2010. Lancet 2012 Dec 15;380(9859):2095-2128. [doi: 10.1016/S0140-6736(12)61728-0] [Medline: 23245604]
3. Miravitlles M, Vogelmeier C, Roche N, Halpin D, Cardoso J, Chuchalin AG, et al. A review of national guidelines for management of COPD in Europe. Eur Respir J 2016 Feb;47(2):625-637 [FREE Full text] [doi: 10.1183/13993003.01170-2015] [Medline: 26797035]
4. Rodriguez-Roisin R. Toward a consensus definition for COPD exacerbations. Chest 2000 May;117(5 Suppl 2):398S-401S. [doi: 10.1378/chest.117.5_suppl_2.398s] [Medline: 10843984]
5. Hoogendoorn M, Feenstra TL, Hoogenveen RT, Al M, Mölken MRV. Association between lung function and exacerbation frequency in patients with COPD. Int J Chron Obstruct Pulmon Dis 2010 Dec 09;5:435-444 [FREE Full text] [doi: 10.2147/COPD.S13826] [Medline: 21191438]
6. Donaldson GC, Seemungal TAR, Bhowmik A, Wedzicha JA. Relationship between exacerbation frequency and lung function decline in chronic obstructive pulmonary disease. Thorax 2002 Oct;57(10):847-852 [FREE Full text] [doi: 10.1136/thorax.57.10.847] [Medline: 12324669]

7.   Donaldson GC, Wedzicha JA. COPD exacerbations .1: Epidemiology. Thorax 2006 Feb;61(2):164-168 [FREE Full text]
     [doi: 10.1136/thx.2005.041806] [Medline: 16443707]

8.   Wilkinson TMA, Donaldson GC, Hurst JR, Seemungal TAR, Wedzicha JA. Early Therapy Improves Outcomes of
     Exacerbations of Chronic Obstructive Pulmonary Disease. Am J Respir Crit Care Med 2004 Jun 15;169(12):1298-1303.
     [doi: 10.1164/rccm.200310-1443oc]

9.   Wedzicha JA, Seemungal TAR. COPD exacerbations: defining their cause and prevention. Lancet 2007 Sep
     01;370(9589):786-796 [FREE Full text] [doi: 10.1016/S0140-6736(07)61382-8] [Medline: 17765528]

10. Adibi A, Sin DD, Safari A, Johnson KM, Aaron SD, FitzGerald JM, et al. The Acute COPD Exacerbation Prediction Tool
     (ACCEPT): a modelling study. Lancet Respir Med 2020 Oct;8(10):1013-1021. [doi: 10.1016/S2213-2600(19)30397-2]
     [Medline: 32178776]

11. Guerra B, Gaveikaite V, Bianchi C, Puhan MA. Prediction models for exacerbations in patients with COPD. Eur Respir
     Rev 2017 Jan;26(143):160061 [FREE Full text] [doi: 10.1183/16000617.0061-2016] [Medline: 28096287]

12. Sobnath DD, Philip N, Kayyali R, Nabhani-Gebara S, Pierscionek B, Vaes AW, et al. Features of a Mobile Support App
     for Patients With Chronic Obstructive Pulmonary Disease: Literature Review and Current Applications. JMIR Mhealth
     Uhealth 2017 Feb 20;5(2):e17 [FREE Full text] [doi: 10.2196/mhealth.4951] [Medline: 28219878]

13. Velardo C, Shah SA, Gibson O, Clifford G, Heneghan C, Rutter H, EDGE COPD Team. Digital health system for personalised
     COPD long-term management. BMC Med Inform Decis Mak 2017 Feb 20;17(1):19 [FREE Full text] [doi:
     10.1186/s12911-017-0414-8] [Medline: 28219430]

14. North M, Bourne S, Green B, Chauhan AJ, Brown T, Winter J, et al. A randomised controlled feasibility trial of E-health
     application supported care vs usual care after exacerbation of COPD: the RESCUE trial. NPJ Digit Med 2020;3:145 [FREE
     Full text] [doi: 10.1038/s41746-020-00347-7] [Medline: 33145441]

15. Jones PW, Harding G, Berry P, Wiklund I, Chen W, Kline Leidy N. Development and first validation of the COPD
     Assessment Test. Eur Respir J 2009 Sep;34(3):648-654 [FREE Full text] [doi: 10.1183/09031936.00102509] [Medline:
     19720809]

16. Dodd JW, Hogg L, Nolan J, Jefford H, Grant A, Lord VM, et al. The COPD assessment test (CAT): response to pulmonary
     rehabilitation. A multicentre, prospective study. Thorax 2011 May;66(5):425-429. [doi: 10.1136/thx.2010.156372] [Medline:
     21398686]

17. Gupta N, Pinto LM, Morogan A, Bourbeau J. The COPD assessment test: a systematic review. Eur Respir J 2014
     Oct;44(4):873-884 [FREE Full text] [doi: 10.1183/09031936.00025214] [Medline: 24993906]

18. Bergstra J, Bardenet R, Bengio Y, Kégl B. Algorithms for hyper-parameter optimization. In Advances in neural information
     processing systems -2554. doi/10.5555/2986459.298 2011;2546:6743. [doi: 10.5555/2986459.2986743]

19. Bergstra J, Yamins D, Cox D. Making a science of model search: Hyperparameter optimization in hundreds of dimensions
     for vision architectures. 2013 Presented at: ICML'13: Proceedings of the 30th International Conference on International
     Conference on Machine Learning; June; Atlanta, GA, USA p. 115-123.

20. Youden WJ. Index for rating diagnostic tests. Cancer 1950 Jan;3(1):32-35. [Medline: 15405679]

21. Varol Y, Ozacar R, Balci G, Usta L, Taymaz Z. Assessing the effectiveness of the COPD Assessment Test (CAT) to evaluate
     COPD severity and exacerbation rates. COPD 2014 Apr;11(2):221-225. [doi: 10.3109/15412555.2013.836169] [Medline:
     24111793]

22. Halpin DM, Miravitlles M, Metzdorf N, Celli B. Impact and prevention of severe exacerbations of COPD: a review of the
     evidence. Int J Chron Obstruct Pulmon Dis 2017;12:2891-2908 [FREE Full text] [doi: 10.2147/COPD.S139470] [Medline:
     29062228]

23. Global strategy for the diagnosis, management and prevention of COPD. Global Initiative for Chronic Obstructive Lung
     Disease (GOLD). 2020. URL: https://goldcopd.org/ [accessed 2020-11-25]

24. Wilkinson TMA, Donaldson GC, Johnston SL, Openshaw PJM, Wedzicha JA. Respiratory syncytial virus, airway
     inflammation, and FEV1 decline in patients with chronic obstructive pulmonary disease. Am J Respir Crit Care Med 2006
     Apr 15;173(8):871-876. [doi: 10.1164/rccm.200509-1489OC] [Medline: 16456141]

25. Tomasic I, Tomasic N, Trobec R, Krpan M, Kelava T. Continuous remote monitoring of COPD patients-justification and
     explanation of the requirements and a survey of the available technologies. Med Biol Eng Comput 2018 Apr;56(4):547-569
     [FREE Full text] [doi: 10.1007/s11517-018-1798-z] [Medline: 29504070]

26. Siafakas N, Vermeire P, Pride N, Paoletti P, Gibson J, Howard P, et al. Optimal assessment and management of chronic
     obstructive pulmonary disease (COPD). The European Respiratory Society Task Force. Eur Respir J 1995 Aug;8(8):1398-1420
     [FREE Full text] [doi: 10.1183/09031936.95.08081398] [Medline: 7489808]

27. Lundberg SM, Nair B, Vavilala MS, Horibe M, Eisses MJ, Adams T, et al. Explainable machine-learning predictions for
     the prevention of hypoxaemia during surgery. Nat Biomed Eng 2018 Oct;2(10):749-760 [FREE Full text] [doi:
     10.1038/s41551-018-0304-0] [Medline: 31001455]

28. Walters EH, Walters J, Wills KE, Robinson A, Wood-Baker R. Clinical diaries in COPD: compliance and utility in predicting
     acute exacerbations. Int J Chron Obstruct Pulmon Dis 2012;7:427-435 [FREE Full text] [doi: 10.2147/COPD.S32222]
     [Medline: 22848156]

29.   Stone AA, Shiffman S. Capturing momentary, self-report data: a proposal for reporting guidelines. Ann Behav Med 2002;24(3):236-243. [doi: 10.1207/S15324796ABM2403_09] [Medline: 12173681]

30.   Bradbury K, Morton K, Band R, van Woezik A, Grist R, McManus RJ, et al. Using the Person-Based Approach to optimise a digital intervention for the management of hypertension. PLoS One 2018;13(5):e0196868 [FREE Full text] [doi: 10.1371/journal.pone.0196868] [Medline: 29723262]

Abbreviations

AUROC: area under the receiver operating characteristic curve
CAT: chronic obstructive pulmonary disease assessment test
COPD: chronic obstructive pulmonary disease

Original Paper

Improving the Prediction of Persistent High Health Care Utilizers: Retrospective Analysis Using Ensemble Methodology

Stephanie N Howson[1], MSc; Michael J McShea[1], MSc; Raghav Ramachandran[1], PhD; Howard S Burkom[1], PhD; Hsien-Yen Chang[2], PhD; Jonathan P Weiner[2], DrPH; Hadi Kharrazi[2], MD, PhD

[1]Applied Physics Laboratory, Johns Hopkins University, Baltimore, MD, United States

[2]Center for Population Health Information Technology, Johns Hopkins School of Public Health, Baltimore, MD, United States

Corresponding Author:
Hadi Kharrazi, MD, PhD
Center for Population Health Information Technology
Johns Hopkins School of Public Health
624 N Broadway
Office 606
Baltimore, MD, 21205-1900
United States
Phone: 1 443 287 8264
Email: kharrazi@jhu.edu

Abstract

Background: A small proportion of high-need patients persistently use the bulk of health care services and incur disproportionate costs. Population health management (PHM) programs often refer to these patients as persistent high utilizers (PHUs). Accurate PHU prediction enables PHM programs to better align scarce health care resources with high-need PHUs while generally improving outcomes. While prior research in PHU prediction has shown promise, traditional regression methods used in these studies have yielded limited accuracy.

Objective: We are seeking to improve PHU predictions with an ensemble approach in a retrospective observational study design using insurance claim records.

Methods: We defined a PHU as a patient with health care costs in the top 20% of all patients for 4 consecutive 6-month periods. We used 2013 claims data to predict PHU status in next 24 months. Our study population included 165,595 patients in the Johns Hopkins Health Care plan, with 8359 (5.1%) patients identified as PHUs in 2014 and 2015. We assessed the performance of several standalone machine learning methods and then an ensemble approach combining multiple models.

Results: The candidate ensemble with complement naïve Bayes and random forest layers produced increased sensitivity and positive predictive value (PPV; 49.0% and 50.3%, respectively) compared to logistic regression (46.8% and 46.1%, respectively).

Conclusions: Our results suggest that ensemble machine learning can improve prediction of care management needs. Improved PPV implies reduced incorrect referral of low-risk patients. With the improved sensitivity/PPV balance of this approach, resources may be directed more efficiently to patients needing them most.

Introduction

Population health management (PHM) programs regularly classify patients by estimated risk of high health care utilization such as hospitalization [1]. The classification process enables PHM programs to allocate their limited resources according to the patients' anticipated needs [1,2]. Higher-risk patient groups,

if identified correctly, can receive effective interventions such as care management program enrollment to reduce utilization and improve outcomes [2]. Additionally, when utilization and costs are successfully contained for high-need patients by proactively preventing undesired outcomes, PHM programs can better allocate the remaining resources to improve the outcomes of other patients [3].

XSL·FO
RenderX

The set of high-risk patients frequently changes over time, with most patients being high-risk for a short term [4,5]. However, some high-risk patients use health care resources persistently for an extended period (eg, more than 24 months) [4-6]. These persistent high utilizer (PHU) patients generally constitute a small segment of the overall patient population but use a considerable proportion of resources in long term [4-6]. Despite the variety of approaches taken to characterize PHUs, such as adjusting for type of utilization, total costs, number of chronic conditions, and other factors, predicting who becomes a PHU has remained an analytical challenge [7-11].

Past studies have applied several analytical approaches to identify and predict PHUs in different patient populations. These approaches range from traditional regression methods (eg, logistic regression) [4-8] to complex machine learning techniques (eg, gradient boosting and neural networks) [9-11]. Nonetheless, due to the small number of PHUs in a patient population (often less than 5%), most studies have suffered from either oversensitive models or excessive false predictions of high utilization [3,5]. Thus, the challenge of achieving simultaneously useful levels of sensitivity and positive predictive value (PPV) in PHU prediction models has limited their application in practice [12].

To address the methodological challenges in predicting PHUs, this study tests an ensemble approach to balance the sensitivity and PPV of PHU forecasting at practical levels. The ensemble approach uses a mix of machine learning methodologies to improve both the sensitivity and PPV of PHU predictions at the same time. Using insurance claims data of a large patient population, this study compares the ensemble approach to single models, a baseline model, and a more advanced predictive model.

Methods

Overall Aims and Definitions

The overall goal of our study was to assess the value of ensemble methodology for achieving required levels of sensitivity and PPV for PHU prediction. Our analysis aimed to provide a methodology to optimize the tradeoff of highly sensitive and highly specific predictive models of PHUs using an ensemble approach.

We defined a PHU as an individual who remained in the top 20% of highest health care costs for 4 consecutive 6-month periods (ie, total of 24 months after the base period) [4]. Health care costs were defined as the sum of costs covered by the insurer and the patient's out-of-pocket costs [4].

Data Source and Preparation

We performed a retrospective analysis of the Johns Hopkins Health Care insurance claims data collected between 2013 and 2015. We applied the Johns Hopkins Adjusted Clinical Groups (ACG) software to the claims data to prepare the data for analysis [13]. We categorized the diagnostic codes into higher-level diagnosis groupings called expanded diagnostic clusters (EDCs), and we grouped medication data into Rx-defined morbidity groups (RxMGs) [4,13]. EDCs and RxMGs have been substantially validated in past studies and are routinely used for risk stratification in practice [4,14].

Study Population

Johns Hopkins Health Care claims data included 207,421 patients with at least 1 record in 2013 and at least 2 years of continuous enrollment between 2013 and 2015 (Figure 1). First, 27,518 patients with missing EDC diagnosis codes were excluded, since EDCs were used to predict PHU status within the population. Second, 14,308 patients with EDC codes indicating pregnancy/newborn status were removed, as the anticipated high utilization incurred by these patients are different from PHUs. The final study population included 165,595 patients (Figure 1).

Figure 1. Selection process of the study population. JHHC: Johns Hopkins Health Care; EDC: expanded diagnostic cluster.
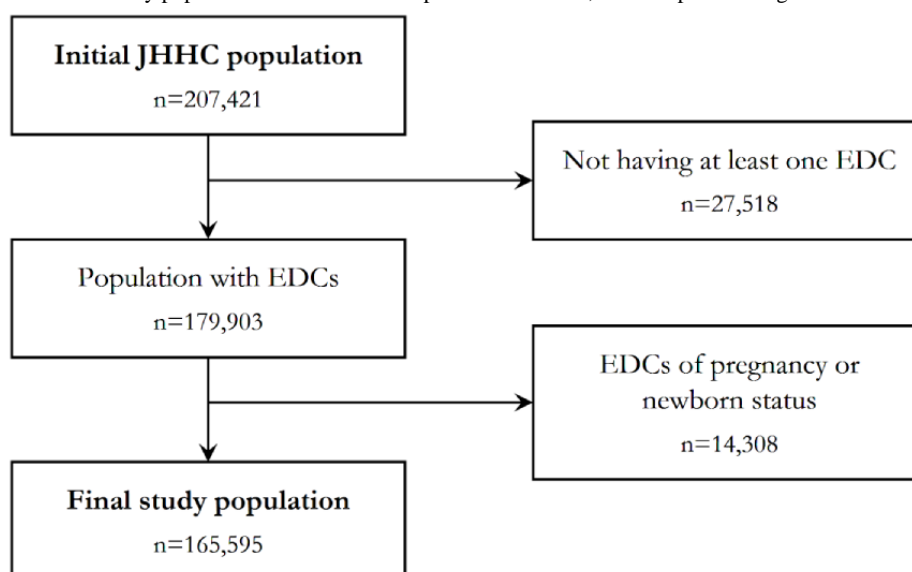
Predictors and Outcome

Predictors (ie, independent variables) included demographics, EDCs, RxMGs, and other health utilization variables (eg, hospitalization) generated by the ACG system. Many of these predictors, including all EDCs and RxMGs, are categorical variables [13,14].

The outcome of interest, a binary variable, was whether a patient became a PHU after the base year (ie, incurred health care costs in the top 20% of all patients over 4 consecutive 6-month periods).

Statistical Approach

Ensemble Methodology

PHUs constitute a small fraction of the patient population, hence producing a large class imbalance (ie, most patients are non-PHUs). A common issue with single model prediction of highly imbalanced classes is compromising PPV in favor of higher sensitivity. For example, a single predictive model of PHUs may result in many false positives (ie, low PPV) if aiming to capture all PHUs (ie, high sensitivity). However, ensemble models provide a unique opportunity to increase both PPV and sensitivity by combining substantially different predictive models. We hypothesized that an ensemble approach can predict PHUs with both a manageable PPV and an optimal sensitivity compared to basic and advanced single model predictions.

We assessed several machine learning algorithms to predict PHU status among the study population. We also evaluated the performance of the ACG system, a comprehensive regression-based risk stratification tool commonly used in PHM

practice [13]. As hypothesized, each of these algorithms yielded average levels of PPV, and we used an ensemble methodology to boost the overall PHU prediction performance.

Ensemble methods take inputs from multiple models and combine the outputs in various ways to strengthen prediction results [15]. In classification problems with imbalanced classes, ensemble methods perform well because multiple models can contribute individual strong features to the overall prediction [16]. Since PHUs make a fraction of the total population, the occurrence of a PHU in the data can be considered an anomaly [4]. Sometimes referred to as anomaly detection, the supervised machine learning problem of classifying PHUs is known as the imbalanced class problem, where the majority class (ie, non-PHUs) is much more prevalent than the minority class (ie, PHUs).

We chose the stacking ensemble model rather than the voting ensemble approach. The stacking ensemble model uses a metaclassifier to aggregate the results, but the voting ensemble model needs user-specified weights to combine the classifiers, hence adding an unpractical step [15]. Thus, for this problem space and our data set, we chose the stacking ensemble. Stacking ensemble methods often use multiple model layers and a final prediction model layer. Each layer makes predictions on the input space given. We also used an additional parameter, feature propagation. This technique allows the passing of both features and predictions through each layer of the ensemble [15]. Figure 2 depicts the overall structure of our ensemble methodology and schematically shows how multiple layers can improve PPV and sensitivity simultaneously (Figure 2).

Figure 2. Stacking ensemble architecture. F&P: feature selection and predictions; PHU: persistent high utilizer; non-PHU: nonpersistent high utilizer.
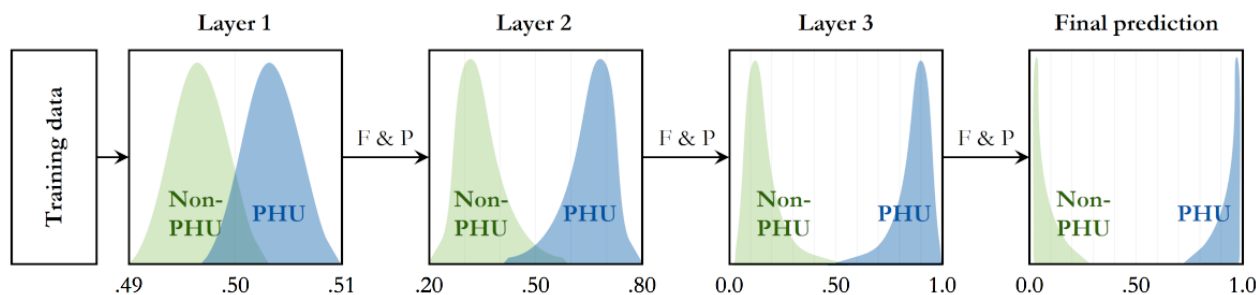
Ensemble Component Model Selection

The models selected as the layers in the ensemble method were chosen using common techniques, namely assessment of common classification algorithms and random search cross-validation for parameter tuning. Typically, machine learning models are assessed for performance and generalizability. Generalizability is difficult to quantify without large unseen data sets available for testing, but a common technique to test for overfitting is k cross-fold validation. This technique tests the machine learning model against many different subsets of data and then calculates an average of all tests. For classifying PHUs, generalizability is fundamentally important because future populations tested through these algorithms will have a large variety of differences, including demographic profiles and medical conditions. Accordingly, we

employed several techniques to tune the performance and generalizability of individual models before constructing the layers of the stacking ensemble [15-17].

First, we incorporated an algorithm known as complement naïve Bayes (CNB), which often produces highly sensitive predictions when classes are imbalanced [18]. The CNB model is derived from standard multinomial naïve Bayes [18]. This model has 3 main parameters, alpha, fit prior, and norm [18]. Alpha is a Laplace smoothing parameter that adjusts the shape and fit of the multinomial distribution. This parameter shifts and forms the training distribution to characterize the multidimensional space of the data. Fit prior refines class identification when only a single class is found in the training set, which can easily occur since PHUs occur infrequently in the data set. Fitting the priors of the classifier ensures that the majority class (ie, non-PHUs)

still has some probability of not occurring, even though no other class is present in the training data. The norm parameter determines whether the training involves a second normalization of weights, an additional measure to bolster the performance on imbalanced class problems like PHU detection. Naïve Bayes models are very easy to train, so a fine-tuned parameter search was performed to find more than 1 robust CNB for use in the stacking ensemble [18].

Second, we integrated a random forest (RF) classifier in the ensemble model. An RF model is a meta-estimator that fits numerous decision tree classifiers on subsets of data features and averages results (ie, polls) to improve performance [19]. Decision trees, and by association RFs, are useful in several applications due to their explainability and ease of training. Decision trees do not require normalization and can accept categorical and numerical variables; however, a shortcoming of decision trees is their difficulty with generalization. Imprecise selection of hyperparameters will make the RF tree overly complex resulting in poor performance when facing unseen patterns [19]. Since RFs are an estimator built by decision trees, many of the parameters are carried over, although additional parameters are available for the sampling and final averaging with the RF [19].

All applicable parameters of an RF were varied through a random cross-validated grid search, but a few most notably contributed to overall performance and generalizability. These parameters include number of estimators, maximum depth, minimum samples to split, minimum samples at leaf, maximum number of features, and class weight. The number of estimators is the count of how many decision trees should be fitted to make up the RF [19]. Increasing the number of estimators typically increases generalizability but must be monitored for computational complexity. Maximum depth fixes the maximum number of levels that each tree can have, which is critical in generalizability [19]. If not set, the tree is continued until each leaf is pure, meaning the tree could learn the pattern of a single person in this population, which is not extensible to unseen populations. Minimum samples to split sets the minimum number of samples at the time of a split, ensuring that each leaf has at least n–1 samples. Minimum samples at leaf is very similar to minimum samples to split but controls samples at the leaf level. In this study, minimum samples at leaf was used to ensure edge cases (ie, unique PHU patterns) were still appropriately populated with training samples. Maximum number of features describes the method used to generate each tree which in certain use cases, taking the square root or log of the total number of features, can increase an RF's performance [19].

Class weight is the most important RF parameter for performance, although setting it can negatively impact generalizability [19]. This parameter adjusts the prior weight on the positive class, which is important for imbalanced classes, and it pushes the decision tree fits to focus more closely on the minority class, making it more robust to edge cases. Since this model was designed to detect PHUs, favoring minority instead of majority class performance was key. Using specific class weights forced the decision trees to allow for a degradation in classifying non-PHUs in favor of an increase in PHU

classification. Two RF models were selected from a random search cross-validation of parameters for use in the stacking ensemble. The final stacking ensemble model integrated the CNB and RF models into one predictive model.

The final ensemble model used an 80/20 split for training and testing of the data. We performed a 5-fold cross-validation on hyperparameter search and recursive feature elimination.

Performance Metrics

Typically, positive and negative class performance are assessed equally using a metric such as F1 score. In this study, as the PHU versus non-PHU classes are unequal and the positive class would constitute an infrequent occurrence, only the positive class metrics were considered key for performance improvement. Therefore, we measured PPV and sensitivity metrics to assess performance of all models (ie, individual models and ensemble model). Both performance metrics describe the classification results for the positive class (ie, PHUs). PPV is the proportion of positive classifications that are truly PHUs. Sensitivity is the proportion of PHUs who were classified as positive.

An important consideration in any machine learning algorithm evaluation is the balance among metrics. A simple way to find an appropriate balance is to change the threshold for classification. Choosing the appropriate threshold can be difficult for health care scenarios due to the risk of incorrect classification for an individual who needs treatment (ie, false negatives). Conversely, classifying too many healthy individuals at risk could overwhelm the resources available for interventions (ie, false positives). To address this issue, we calculated and then plotted sensitivity and PPV for 50 trials at thresholds spaced evenly .05 apart. We then calculated the discrimination threshold for the ensemble model to choose the optimal threshold of the PPV versus sensitivity metrics.

Finally, we compared the PPV and sensitivity of select individual models, which achieved at least 40% performance in both metrics, with the ensemble methodology. The individual models included a logistic regression, the Johns Hopkins ACG model (out-of-box and with no further training) [13], and a standalone RF model. The ensemble model included a stacking ensemble with multiple layers combining CNB and RF models.

All analyses, including descriptive analysis, individual modeling, and ensemble approach, were performed in R (version 3.5.1, R Foundation for Statistical Computing). We used Python pandas and scikit-learn for all modeling pipeline efforts (eg, data cleaning, filtering, hyperparameter search, feature selection, and RF model). We used Python ML Ensemble for the ensemble model [20]. We used Python Yellowbrick library to visualize the classification threshold of sensitivity versus positive predictive values. We used the Johns Hopkins ACG system to produce the ACG output and measure the ACG model's performance [13].

Results

Descriptive Analyses

The study population comprised 165,595 unique patients including 8359 (5.1%) PHUs (Table 1). The PHU population's average age was more than twice that of the non-PHU population (38.51 years vs 18.79 years). PHUs included fewer males (2735/8359, 32.7%) than non-PHUs (69,683/155,862, 44.7%). As expected, PHUs had more utilization than non-PHUs (1567/8359, 18.7% vs 3891/155,862, 2.5% for inpatient visits and 8332/8359, 99.7% vs 152,199/155,862, 97.3% for outpatient visits, respectively).

Table 1. Specification of the study populations (n=165,595).

	Overall study population (n=165,595)	Non-PHU[a] population (n=155,862)	PHU population (n=8359)
Age (years), mean (SD)	19.85 (17.45)	18.79 (16.82)	38.51 (18.01)
0-17, n (%)	101,264 (61.2)	99,352 (63.7)	1459 (17.5)
18-64, n (%)	63,260 (38.2)	55,666 (35.7)	6730 (80.5)
65+, n (%)	1037 (0.6)	844 (0.5)	170 (2.0)
Sex (male), n (%)	72,974 (44.1)	69,683 (44.7)	2735 (32.7)
Race, n (%)			
White	41,492 (25.1)	38,762 (24.9)	2457 (29.4)
Black	54,207 (32.7)	50,993 (32.7)	2879 (34.4)
Other[b]	149 (0.1)	143 (0.1)	6 (<0.1)
Missing[c]	69,747 (42.1)	65,964 (42.3)	3017 (36.1)
Inpatient visits, n (%)			
0	160,035 (96.6)	151,971 (97.5)	6792 (81.3)
1-5	5430 (3.3)	3866 (2.5)	1500 (17.9)
6-10	77 (<0.1)	20 (<0.1)	54 (0.6)
11+	19 (<0.1)	5 (<0.1)	13 (0.2)
Outpatient visits, n (%)			
0	3720 (2.2)	3663 (2.4)	27 (0.3)
1-5	96,122 (58.0)	94,138 (60.4)	1234 (14.8)
6-10	33,996 (20.5)	32,317 (20.7)	1428 (17.1)
11+	31,723 (19.2)	25,744 (16.5)	5670 (67.8)

[a]PHU: persistent high utilizer.

[b]Members of known race/ethnicity not equal to Asian, Hispanic, White, or Black.

[c]Members with empty values for race.

Ensemble Model

After tuning the ensemble layers, the best-performing ensemble model included 3 input layers and 1 prediction layer. The final ensemble model included 2 input layers of CNB and 1 layer of an RF model. The prediction layer was an RF model. The model included the following variables: race (ie, Black, White, other), age (as of 2013), sex, days of inpatient hospitalization in 2013, emergency department visit count in 2013, psychotherapy services in 2013, outpatient visit count in 2013, all-cause inpatient hospitalization count in 2013, frailty flag for older adults, 87 most frequent Johns Hopkins ACG diagnostic comorbidities (ie, EDCs [13]), all Johns Hopkins ACG medication grouping (ie, RxMGs [13]), and ACG-derived care coordination risk scores [13] (ie, likely coordination issue, possible coordination issue, unlikely coordination issue). These variables are generated by and included in the John Hopkins ACG risk stratification models, which are widely used for PHM efforts [13]. The stacking ensemble had full feature propagation throughout the layers to allow each model access to all data attributes while gaining classification scores from previous layers. The most performant models were selected for use in the stacking ensemble.

Model Performance Evaluation

Figure 3 depicts the discrimination threshold plot for a sample decision tree of the ensemble model. The plot conveys the importance of the threshold choice and depicts the tradeoff between PPV and sensitivity. As shown in the figure, patients A and B, both of whom are PHUs, will be identified differently by the model depending on the chosen threshold between PPV and sensitivity. By testing the trained model on these 2 patients, a risk score is generated for each. These risk scores can be compared to any classification threshold. Depending on which side of the threshold the risk scores lie, the model classified whether patient A, B, or both are PHUs or non-PHUs.

Figure 3. Classification threshold of sensitivity versus positive predictive value (PPV): patient A: incorrectly classified as normal (risk score=82%) and patient B: correctly classified as a persistent high utilizer (risk score=97%).
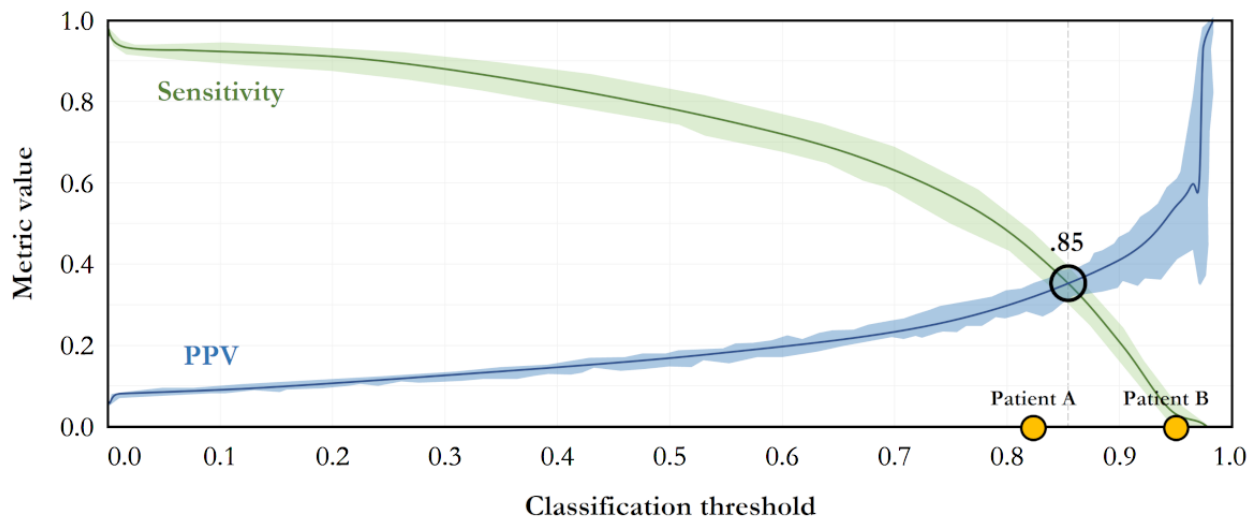
The central line in Figure 3 represents the median value for each metric, and the bands represent the variability from the 10th to 90th percentiles. Two important observations about the threshold plot are (1) the typical classification threshold of .50 is not ideal probably due to the imbalanced classes and (2) equally weighting sensitivity and PPV at a threshold of .85 may not be appropriate to classify enough PHUs correctly. Patients A and B in Figure 3 have different classification outcomes and therefore interventions due in part to an arbitrary threshold.

To replicate the same level of optimality across all models, we used the 95th percentile threshold limit for each model. The absolute cutoff points were slightly different across models with ensemble having an absolute cutoff threshold of .258, RF .224, logistic regression .230, and the ACG model a cutoff of .226. Negative predictive value (NPV) and specificity were also assessed, but performance in these metrics was high (ie, averaging 97% and 99% for NPV and specificity, respectively) and did not vary significantly between models due to the large size and variability of the negative class (ie, non-PHUs).

Performance Comparison

The stacking ensemble method achieved a sensitivity of 49.0% and PPV of 50.3%. The ensemble model resulted in a 5%+ increase in both PPV and sensitivity for predicting PHUs over other individual methods such as logistic regression, RF model, and the ACG model (Table 2). As shown in Table 2, the individual RF was the highest performing nonensemble technique. Table 2 also includes the optimal parameters used in the stacking ensemble (eg, CNB and RF parameters such as alpha, maximum depth, and minimum sample splits). The final ensemble model also produced an NPV of 97.4%, specificity of 97.3%, and F1 of 49.1% for PHUs and 97.4% for non-PHUs (not shown in Table 2). The area under the curve of the ensemble model reached .921; however, comparison of areas under the curve between models was considered not valuable due to the large imbalance of PHUs versus non-PHUs, hence limiting the performance measure comparison to PPV and sensitivity of the models.

Table 2. Model fit statistics for predicting persistent high utilizer status.

Model	Parameter tuning	Sensitivity, %	PPV[a], %
Stacking ensemble	CNB1 =.70, fit prior, norm	49.0	50.3
Layer 1: CNB[b]	CNB2 =.15, fit prior		
Layer 2: CNB	RF1		
Layer 3: RF[c]	• 200 estimators		
Prediction layer: RF	• 400 max[d] depth		
Feature propagation	• 5 min[e] samples split		
	• 0.01% min samples		
	Leaf		
	• auto max features		
	• class weight=0.842		
	RF2		
	• 100 estimators		
	• 350 max depth		
	• 2 min samples split		
	• 0.01% min samples		
	Leaf		
	• class weight=1.0		
RF	• 300 estimators	48.4	47.2
	• 500 max depth		
	• 20 min samples split		
	• 0.01% min samples leaf		
JHU-ACG[f]	ACG[g] system probability of PHU[h]	44.7	44.1
Logistic regression	Based on 241 parameters (ie, diagnoses and medications)	46.8	46.1

[a]PPV: positive predictive value.

[b]CNB: complement naïve Bayes.

[c]RF: random forest.

[d]max: maximum.

[e]min: minimum.

[f]JHU-ACG: ACG predictive model with no local tuning.

[g]ACG: adjusted clinical group.

[h]PHU: persistent high utilizer.

Discussion

Principal Findings

Persistent high utilizers (PHUs) are defined as patients who consistently stay in the highest deciles of health care costs or utilization across multiple years [4-12]. Risk stratification efforts strive to better identify and manage PHUs so that scarce health care resources can be better allocated. Nonetheless, predicting who becomes a PHU is often challenging, partly because PHUs are uncommon [4,6,9-11]. Past studies have attempted to improve the prediction of PHUs in various populations; however, those predictions have either suffered from high false negative/positive rates or have been limited in scope [4,6,9-11]. In this study, to address the methodological complexity in predicting PHUs, we evaluated the benefit of an ensemble approach to balance the sensitivity and specificity of predicting PHUs.

Our results show that ensemble methodology can be effectively used to improve both sensitivity and PPV of predicting PHUs.

The ensemble model developed in this study included 2 layers of CNB and 1 prediction layer of RF, which can be converged rather quickly. We achieved a sensitivity and PPV of 49.0% and 50.3%, respectively, using the ensemble model. In comparison to the best alternative performing model, which was the standalone RF, the ensemble model improved the sensitivity by 0.6 and PPV by 3.1 absolute percentage points, which represents a 1.2% and 6.6% relative improvement in sensitivity and PPV, respectively. Moreover, standalone RF models are prone to overfitting and often lack generalizability to other populations. The ensemble model was also superior compared to traditional logistic regression and the more established (ACG) models [13]. The ensemble model improved the sensitivity and PPV of predicting PHUs by 2.2 and 4.2 absolute percentage points (ie, 4.7% and 9.1% relative improvement) compared to the traditional logistic regression and by 4.3 and 6.2 absolute percentage points (ie, 9.6% and 14.1% relative improvement) when compared to the ACG model [13].

Several studies have examined the use of traditional methods in predicting PHUs; however, models developed in these studies have often generated low PPV rates or showed limited generalizability. For example, in a study of an employer-based health plan, using commercial claims data, a logistic regression model achieved a sensitivity of 80% but PPV of 19% to predict PHUs among the health plan enrollees [6]. In another study aiming to predict PHUs, using diagnostic and medication information extracted from claims data, a regression model achieved a sensitivity of 46.7% and PPV of 57.2%; however, the study population was limited to patients aged 18 to 62 years, hence limiting generalizability to other populations [4]. Several studies have used regression models to control for underlying demographic and clinical variables and measure the residual differences such as cost, behavioral health, and social determinants of health variables between PHU and non-PHU populations [7,8]. These studies, however, have not published the performance of these regression models in predicting PHUs.

A few studies have assessed the value of machine learning methods in predicting PHUs. In a study of a statewide Medicaid population, demographics, diagnostics, and medication information were used to predict costs associated with PHUs. The study compared multiple models including linear regression, regularized regression, gradient boosting machine, and recurrent neural networks, but the study did not generate comparable predictive measures as these models did not predict PHU status [9]. Another study applied penalized regression, support vector machine, and extreme gradient boosting against claims data to predict PHUs among patients from an academic medical center. The study achieved high sensitivity rates ranging from 72.7% to 78.7%; however, the (recalculated) PPV ranged from 18.6% to 19.8% [10]. Among the machine learning studies targeting PHUs, only one study compared an ensemble methodology (using RFs) to other methods (eg, linear regression, decision tree regression) [11]. This study, however, predicted cost of PHUs and was limited to patients with schizophrenia, hence limiting its generalizability to the broader population of patients.

Despite the promising findings of past studies in predicting PHUs, their results cannot be accurately compared to our ensemble model as each study used a slightly different definition of PHU. Some studies have defined PHUs as patients in the top 5% of cost over 2 years [4], while other studies have set the bar at 10% or 20% of cost over longer periods of time [6,7]. Future research should attempt to harmonize the definition of PHUs to make the comparison of PHU populations across different populations and health plans feasible. Additionally, harmonization of the PHU definition can facilitate the performance measurement and comparison of PHU predictive models across different health care settings.

Balancing the sensitivity and PPV of PHU predictions is key in operationalizing such models in PHM efforts. Indeed, given the infrequency of PHUs in the total population of patients, a balanced sensitivity and PPV ratio will play an important role in the management of limited resources for PHUs. In our study, the improvement of model performance compared to the traditional models corresponds to approximately 84 additional PHUs being classified correctly in the test set of 1672 true PHUs. These 84 patients would not have been reviewed for potential proactive interventions by a care manager if tested by a traditional method.

In this study, we chose to report classification performance at the balanced precision and recall scores (50/50) to highlight optimal performance in both metrics simultaneously. In specific PHM use cases, it may be desirable to select a lower classification threshold and more patients for care or intervention consideration, even if their individual risk score is lower. In large-scale PHM use cases, cost of considering many patients may be too high and a higher classification threshold is to be selected to only manage the most at-risk patients. Hence, individual population health programs may chose different balances of precision versus recall for models predicting PHUs.

Our study showed that machine learning has a performance advantage over traditional statistical models. Ultimately, improved performance will come from more advanced ensemble methods coupled with continually improving robustness of feature analysis, which together are the keys to significantly increased performance. Model performance could benefit from subpopulation training by reducing the large and variable parameter space for classification. Thus, developing custom groupings of clinical features associate with PHU patients (versus non-PHUs) can potentially advance predictive models of PHUs. For example, clinical groupings identified by unsupervised machine learning techniques (such as latent class analysis) has shown value in improving predictive models of PHUs [21].

Value-based health care providers are increasingly using risk stratification tools to manage their patient populations [22]. Providers often use local electronic health records (EHRs) instead of insurance claims to risk stratify patients and predict PHUs [23-25]. Although advances have been made in using unique EHR data to improve risk prediction using prescription data [26-28], vital signs [29,30], laboratory results [31], and free-text analysis [32,33], quality of EHR data remains a major challenge in this process [34]. Using machine learning models, such as the ensemble models, can potentially help providers address some of these deficiencies and improve the prediction of PHUs using EHR data [35,36]. Future studies should investigate the usability of machine learning models in enhancing EHR-based PHU predictions and its implication on improving the wider population-level health outcomes [37].

Limitations

Our study has several limitations. First, the results of our ensemble approach and the improvement of the PHU prediction may not generalize to other populations (eg, older adults), different settings (eg, inpatient only), or alternative data sources (eg, EHRs). Future research should explore the use of ensemble methodology in new populations and settings using alternate data sources. Second, the current definition of PHU may not be consistent with the operational definition in all PHM. We used a specific definition for PHU (ie, percentile of cost and time period), but that definition may not fit all populations. The risk stratification research community should harmonize the definition of PHU so predictive models of PHUs can be compared accurately to increase their generalizability. Third, we only used demographics, diagnosis, and medications in our

prediction models. Past research has shown the value of social determinants of health in improving the prediction of health care utilization [38-42]. Future research should investigate the value of the ensemble model in improving predictive models of PHU that incorporate social data. Finally, the ensemble methodology uses an approach that complicates the explanation of a prediction, and thus the operational use of such models in clinical and PHM settings should be further studied.

Conclusion

A small segment of the patient population uses most of the health care services over extended periods. We used an ensemble model, a machine learning approach that combines multiple modeling techniques, to simultaneously improve the sensitivity and PPV of predicting PHUs using claims data. Future studies should investigate the value of machine learning techniques in predicting PHUs in other health care settings with potentially different underlying populations and different data sources (eg, EHR data).

Authors' Contributions

HK and MJM codirected the research project. SNH analyzed the data. HYC provided analytical insight and calculated claims costs. HK, MJM, HSB, RR, and JPW reviewed and interpreted the results. HK, SNH, and MJM drafted the manuscript. All authors reviewed and contributed to the final manuscript. HK prepared the manuscript for submission.

Conflicts of Interest

None declared.

References

1. Iezzoni LI. Risk Adjustment for Measuring Health Care Outcomes, Fourth Edition. New York: Health Administration Press; 2012.

2. Kharrazi H, Gamache R, Weiner J. Role of informatics in bridging public and population health. In: Magnuson J, Dixon B, editors. Public Health Informatics and Information Systems. London: Springer; 2020.

3. Lee NS, Whitman N, Vakharia N, Ph DBT, Rothberg MB. High-cost patients: hot-spotters don't explain the half of it. J Gen Intern Med 2017 Jan;32(1):28-34 [FREE Full text] [doi: 10.1007/s11606-016-3790-3] [Medline: 27480529]

4. Chang H, Boyd CM, Leff B, Lemke KW, Bodycombe DP, Weiner JP. Identifying consistent high-cost users in a health plan: comparison of alternative prediction models. Med Care 2016 Sep;54(9):852-859. [doi: 10.1097/MLR.0000000000000566] [Medline: 27326548]

5. Guilcher SJT, Bronskill SE, Guan J, Wodchis WP. Who are the high-cost users? A method for person-centred attribution of health care spending. PLoS One 2016;11(3):e0149179 [FREE Full text] [doi: 10.1371/journal.pone.0149179] [Medline: 26937955]

6. Hwang W, LaClair M, Camacho F, Paz H. Persistent high utilization in a privately insured population. Am J Manag Care 2015 Apr;21(4):309-316 [FREE Full text] [Medline: 26014469]

7. Yoon J, Chee CP, Su P, Almenoff P, Zulman DM, Wagner TH. Persistence of high health care costs among VA patients. Health Serv Res 2018 Oct;53(5):3898-3916 [FREE Full text] [doi: 10.1111/1475-6773.12989] [Medline: 29862504]

8. Sterling S, Chi F, Weisner C, Grant R, Pruzansky A, Bui S, et al. Association of behavioral health factors and social determinants of health with high and persistently high healthcare costs. Prev Med Rep 2018 Sep;11:154-159 [FREE Full text] [doi: 10.1016/j.pmedr.2018.06.017] [Medline: 30003015]

9. Yang C, Delcher C, Shenkman E, Ranka S. Machine learning approaches for predicting high cost high need patient expenditures in health care. Biomed Eng Online 2018 Nov 20;17(Suppl 1):131 [FREE Full text] [doi: 10.1186/s12938-018-0568-3] [Medline: 30458798]

10. Ng SHX, Rahman N, Ang IYH, Sridharan S, Ramachandran S, Wang DD, et al. Characterising and predicting persistent high-cost utilisers in healthcare: a retrospective cohort study in Singapore. BMJ Open 2020 Jan 06;10(1):e031622 [FREE Full text] [doi: 10.1136/bmjopen-2019-031622] [Medline: 31911514]

11. Wang Y, Iyengar V, Hu J, Kho D, Falconer E, Docherty JP, et al. Predicting future high-cost schizophrenia patients using high-dimensional administrative data. Front Psychiatry 2017;8:114 [FREE Full text] [doi: 10.3389/fpsyt.2017.00114] [Medline: 28713293]

12. Wodchis WP, Austin PC, Henry DA. A 3-year study of high-cost users of health care. CMAJ 2016 Feb 16;188(3):182-188 [FREE Full text] [doi: 10.1503/cmaj.150064] [Medline: 26755672]

13. Johns Hopkins ACGs System, version 12. Johns Hopkins School of Public Health. 2019. URL: https://www.hopkinsacg.org/ [accessed 2022-02-07]

14. Weiner JP, Starfield BH, Steinwachs DM, Mumford LM. Development and application of a population-oriented measure of ambulatory care case-mix. Med Care 1991 May;29(5):452-472. [doi: 10.1097/00005650-199105000-00006] [Medline: 1902278]

15. Zhi-Hua Z. Ensemble Methods: Foundations and Algorithms, 1st Edition. New York: Chapman and Hall/CRC; 2012.

16. Chen Z, Duan J, Kang L, Qiu G. Class-imbalanced deep learning via a class-balanced ensemble. IEEE Trans Neural Netw Learn Syst 2021 Apr 26:1. [doi: 10.1109/TNNLS.2021.3071122] [Medline: 33900923]

17. Yu K, Xie X. Predicting hospital readmission: a joint ensemble-learning model. IEEE J Biomed Health Inform 2020 Feb;24(2):447-456. [doi: 10.1109/JBHI.2019.2938995] [Medline: 31484143]

18. Rennie J, Shih L, Teevan J, Karger D. Tackling the poor assumptions of naive Bayes text classifiers. Proc 20th Int Conf Mach Learn 2003;3:616-623.

19. Breiman L. Random forests. Machine Learning 2001;45:5-32. [doi: 10.1023/A:1010933404324]

20. Flennerhag S. ML-Ensemble: high performance ensemble learning in Python. URL: http://ml-ensemble.com/ [accessed 2022-02-09]

21. Ramachandran R, McShea MJ, Howson SN, Burkom HS, Chang H, Weiner JP, et al. Assessing the value of unsupervised clustering in predicting persistent high health care utilizers: retrospective analysis of insurance claims data. JMIR Med Inform 2021 Nov 25;9(11):e31442 [FREE Full text] [doi: 10.2196/31442] [Medline: 34592712]

22. Pandya CJ, Chang H, Kharrazi H. Electronic health record-based risk stratification: a potential key ingredient to achieving value-based care. Popul Health Manag 2021 Jun 14;24(6):654-656. [doi: 10.1089/pop.2021.0131] [Medline: 34129398]

23. Kharrazi H, Chi W, Chang H, Richards TM, Gallagher JM, Knudson SM, et al. Comparing population-based risk-stratification model performance using demographic, diagnosis and medication data extracted from outpatient electronic health records versus administrative claims. Med Care 2017 Aug;55(8):789-796. [doi: 10.1097/MLR.0000000000000754] [Medline: 28598890]

24. Kharrazi H, Weiner JP. A practical comparison between the predictive power of population-based risk stratification models using data from electronic health records versus administrative claims: setting a baseline for future EHR-derived risk stratification models. Med Care 2018 Dec;56(2):202-203. [doi: 10.1097/MLR.0000000000000849] [Medline: 29200132]

25. Kharrazi H, Gonzalez CP, Lowe KB, Huerta TR, Ford EW. Forecasting the maturation of electronic health record functions among US hospitals: retrospective analysis and predictive model. J Med Internet Res 2018 Aug 07;20(8):e10458 [FREE Full text] [doi: 10.2196/10458] [Medline: 30087090]

26. Chang H, Richards TM, Shermock KM, Elder Dalpoas S, J Kan H, Alexander GC, et al. Evaluating the impact of prescription fill rates on risk stratification model performance. Med Care 2017 Dec;55(12):1052-1060. [doi: 10.1097/MLR.0000000000000825] [Medline: 29036011]

27. Kharrazi H, Ma X, Chang H, Richards TM, Jung C. Comparing the predictive effects of patient medication adherence indices in electronic health record and claims-based risk stratification models. Popul Health Manag 2021 Oct;24(5):601-609. [doi: 10.1089/pop.2020.0306] [Medline: 33544044]

28. Chang H, Kan HJ, Shermock KM, Alexander GC, Weiner JP, Kharrazi H. Integrating e-prescribing and pharmacy claims data for predictive modeling: comparing costs and utilization of health plan members who fill their initial medications with those who do not. J Manag Care Spec Pharm 2020 Oct;26(10):1282-1290. [doi: 10.18553/jmcp.2020.26.10.1282] [Medline: 32996394]

29. Kharrazi H, Chang H, Heins SE, Weiner JP, Gudzune KA. Assessing the impact of body mass index information on the performance of risk adjustment models in predicting health care costs and utilization. Med Care 2018 Dec;56(12):1042-1050. [doi: 10.1097/MLR.0000000000001001] [Medline: 30339574]

30. Kharrazi H, Chang H, Weiner JP, Gudzune KA. Assessing the added value of blood pressure information derived from electronic health records in predicting health care cost and utilization. Popul Health Manag 2021 Nov 29:250. [doi: 10.1089/pop.2021.0250] [Medline: 34847729]

31. Lemke KW, Gudzune KA, Kharrazi H, Weiner JP. Assessing markers from ambulatory laboratory tests for predicting high-risk patients. Am J Manag Care 2018 Jun 01;24(6):e190-e195 [FREE Full text] [Medline: 29939509]

32. Kan HJ, Kharrazi H, Leff B, Boyd C, Davison A, Chang H, et al. Defining and assessing geriatric risk factors and associated health care utilization among older adults using claims and electronic health records. Med Care 2018 Dec;56(3):233-239. [doi: 10.1097/MLR.0000000000000865] [Medline: 29438193]

33. Kharrazi H, Anzaldi LJ, Hernandez L, Davison A, Boyd CM, Leff B, et al. The value of unstructured electronic health record data in geriatric syndrome case identification. J Am Geriatr Soc 2018 Aug;66(8):1499-1507. [doi: 10.1111/jgs.15411] [Medline: 29972595]

34. Kharrazi H, Wang C, Scharfstein D. Prospective EHR-based clinical trials: the challenge of missing data. J Gen Intern Med 2014 Jul;29(7):976-978 [FREE Full text] [doi: 10.1007/s11606-014-2883-0] [Medline: 24839057]

35. Ng SH, Rahman N, Ang IYH, Sridharan S, Ramachandran S, Wang DD, et al. Characterization of high healthcare utilizer groups using administrative data from an electronic medical record database. BMC Health Serv Res 2019 Jul 05;19(1):452 [FREE Full text] [doi: 10.1186/s12913-019-4239-2] [Medline: 31277649]

36. Kan HJ, Kharrazi H, Chang H, Bodycombe D, Lemke K, Weiner JP. Exploring the use of machine learning for risk adjustment: a comparison of standard and penalized linear regression models in predicting health care costs in older adults. PLoS One 2019;14(3):e0213258 [FREE Full text] [doi: 10.1371/journal.pone.0213258] [Medline: 30840682]

37. Gamache R, Kharrazi H, Weiner JP. Public and population health informatics: the bridging of big data to benefit communities. Yearb Med Inform 2018 Aug;27(1):199-206 [FREE Full text] [doi: 10.1055/s-0038-1667081] [Medline: 30157524]

38. Hatef E, Ma X, Rouhizadeh M, Singh G, Weiner JP, Kharrazi H. Assessing the impact of social needs and social determinants of health on health care utilization: using patient- and community-level data. Popul Health Manag 2021 Apr;24(2):222-230. [doi: 10.1089/pop.2020.0043] [Medline: 32598228]

39. Chang H, Hatef E, Ma X, Weiner JP, Kharrazi H. Impact of area deprivation index on the performance of claims-based risk-adjustment models in predicting health care costs and utilization. Popul Health Manag 2021 Jun;24(3):403-411. [doi: 10.1089/pop.2020.0135] [Medline: 33434448]

40. Hatef E, Kharrazi H, Nelson K, Sylling P, Ma X, Lasser EC, et al. The association between neighborhood socioeconomic and housing characteristics with hospitalization: results of a national study of veterans. J Am Board Fam Med 2019;32(6):890-903 [FREE Full text] [doi: 10.3122/jabfm.2019.06.190138] [Medline: 31704758]

41. Tan M, Hatef E, Taghipour D, Vyas K, Kharrazi H, Gottlieb L, et al. Including social and behavioral determinants in predictive models: trends, challenges, and opportunities. JMIR Med Inform 2020 Sep 08;8(9):e18084 [FREE Full text] [doi: 10.2196/18084] [Medline: 32897240]

42. Vest JR, Adler-Milstein J, Gottlieb LM, Bian J, Campion TR, Cohen GR, et al. Assessment of structured data elements for social risk factors. Am J Manag Care 2022 Jan 01;28(1):e14-e23 [FREE Full text] [doi: 10.37765/ajmc.2022.88816] [Medline: 35049262]

Abbreviations

ACG: Adjusted Clinical Group
CNB: complement naïve Bayes
EDC: expanded diagnostic cluster
EHR: electronic health record
NPV: negative predictive value
PHM: population health management
PHU: persistent high utilizer
PPV: positive predictive value
RF: random forest
RxMG: Rx-defined morbidity group

Original Paper

Machine Learning Models for Predicting Influential Factors of Early Outcomes in Acute Ischemic Stroke: Registry-Based Study

Po-Yuan Su[1*], MSc; Yi-Chia Wei[2,3,4*], MD; Hao Luo[1], MSc; Chi-Hung Liu[5,6], MD, MSc; Wen-Yi Huang[2,6], MD, PhD; Kuan-Fu Chen[7,8], MD, PhD; Ching-Po Lin[3], PhD; Hung-Yu Wei[1*], PhD; Tsong-Hai Lee[5,6*], MD

[1]Department of Electrical Engineering, National Taiwan University, Taipei, Taiwan

[2]Department of Neurology, Chang Gung Memorial Hospital, Keelung, Taiwan

[3]Institute of Neuroscience, National Yang Ming Chiao Tung University, Taipei, Taiwan

[4]Community Medicine Research Center, Chang Gung Memorial Hospital, Keelung, Taiwan

[5]Department of Neurology, Linkou Chang Gung Memorial Hospital, Taoyuan City, Taiwan

[6]College of Medicine, Chang Gung University, Taoyuan, Taiwan

[7]Clinical Informatics and Medical Statistics Research Center, Chung Gung University, Taoyuan, Taiwan

[8]Department of Emergency, Chang Gung Memorial Hospital, Keelung, Taiwan

[*]these authors contributed equally

Corresponding Author:
Hung-Yu Wei, PhD
Department of Electrical Engineering
National Taiwan University
EE2-238, No. 1, Sec. 4, Roosevelt Rd.
Taipei, 106
Taiwan
Phone: 886 2 33663688
Email: hywei@ntu.edu.tw

Abstract

Background: Timely and accurate outcome prediction plays a vital role in guiding clinical decisions on acute ischemic stroke. Early condition deterioration and severity after the acute stage are determinants for long-term outcomes. Therefore, predicting early outcomes is crucial in acute stroke management. However, interpreting the predictions and transforming them into clinically explainable concepts are as important as the predictions themselves.

Objective: This work focused on machine learning model analysis in predicting the early outcomes of ischemic stroke and used model explanation skills in interpreting the results.

Methods: Acute ischemic stroke patients registered on the Stroke Registry of the Chang Gung Healthcare System (SRICHS) in 2009 were enrolled for machine learning predictions of the two primary outcomes: modified Rankin Scale (mRS) at hospital discharge and in-hospital deterioration. We compared 4 machine learning models, namely support vector machine (SVM), random forest (RF), light gradient boosting machine (LGBM), and deep neural network (DNN), with the area under the curve (AUC) of the receiver operating characteristic curve. Further, 3 resampling methods, random under sampling (RUS), random over sampling, and the synthetic minority over-sampling technique, dealt with the imbalanced data. The models were explained based on the ranking of feature importance and the SHapley Additive exPlanations (SHAP).

Results: RF performed well in both outcomes (discharge mRS: mean AUC 0.829, SD 0.018; in-hospital deterioration: mean AUC 0.710, SD 0.023 on original data and 0.728, SD 0.036 on resampled data with RUS for imbalanced data). In addition, DNN outperformed other models in predicting in-hospital deterioration on data without resampling (mean AUC 0.732, SD 0.064). In general, resampling contributed to the limited improvement of model performance in predicting in-hospital deterioration using imbalanced data. The features obtained from the National Institutes of Health Stroke Scale (NIHSS), white blood cell differential counts, and age were the key features for predicting discharge mRS. In contrast, the NIHSS total score, initial blood pressure, having diabetes mellitus, and features from hemograms were the most important features in predicting in-hospital deterioration. The SHAP summary described the impacts of the feature values on each outcome prediction.

Conclusions: Machine learning models are feasible in predicting early stroke outcomes. An enriched feature bank could improve model performance. Initial neurological levels and age determined the activity independence at hospital discharge. In addition,

physiological and laboratory surveillance aided in predicting in-hospital deterioration. The use of the SHAP explanatory method successfully transformed machine learning predictions into clinically meaningful results.

Introduction

Cerebrovascular disease ranks as the second leading cause of death in the United States and the third cause of disability-adjusted life years (DALYs) globally in 2010 [1]. Ischemic stroke shows higher incidence and prevalence than hemorrhagic stroke. Ischemic stroke survivors commonly have disabilities and substantial function loss that significantly affect their quality of life. Outcome prediction provides a reference for doctors to select rehabilitation strategies and provides patients with decent expectations in the future [2,3]. Several studies have focused on stroke prediction by indicators collected at emergency room (ER) or first at ward admissions [4,5]. In the past, scores such as the Acute Stroke Registry and Analysis of Lausanne (ASTRAL), DRAGON, and SEDAN were used for stroke outcome prediction and proved more accurate than physicians [6]. Over the past few years, most research on stroke prediction has emphasized the use of machine learning, which achieves better performance in predicting stroke outcomes [7]. Recent studies on stroke prediction can be classified into three categories: studies investigating longitudinal data such as health insurance databases for predicting the probability of stroke occurrence, studies predicting recovery in a specific time using numerical data, and studies applying novel machine learning models such as computer vision models [8] or natural language processing models for more accurate diagnosis [9,10].

This work aimed to predict early outcomes using numerical data and applying novel machine learning models, including neural networks and gradient boosting machines for predictions. The specific goals were to predict the modified Rankin Scale (mRS) score at hospital discharge and deterioration during admission. We focused on model performance comparison, ranking of feature importance, and explanation of model predictions. We leveraged the SHapley Additive exPlanations (SHAP) to depict the stroke prediction models and guarantee that the models predict with a solid basis. For imbalanced prediction targets, preprocessing was performed with different resampling methods to balance the data set before model performance comparisons.

Methods

Database

Patient data were collected from January 1 to December 31, 2009, by the Stroke Registry in Chang Gung Healthcare System (SRICHS) [11]. SRICHS is a stroke registry system that prospectively collected patients' clinical information with the ICD 9 diagnostic code 430-437 for acute ischemic and hemorrhagic stroke since 2007. The registry data were anonymized and deidentified before analysis. The data automatically downloaded from the hospital information system included demographic information, laboratory tests, examination reports, and structured information from the electronic medical chart. The data cleaning process included 2 steps. First, the data without the initial blood pressure recordings at admission, mRS at ward admission and discharge, and laboratory hemograms were removed. Second, the data with out-of-range scores on the National Institutes of Health Stroke Scale (NIHSS) were removed, which were attributed to misrecording. The Institutional Review Board of Chang Gung Memorial Hospital approved this study (no. 103-1519C, no. 201900732B0, and no. 201801763A3).

Outcome Measurements

The primary target variable was the mRS at discharge [12]. To turn the prediction issue into a binary classification problem and compare our results directly with the existing methods, we discretized the mRS into two classes: good outcomes defined by mRS 0-2 and poor outcomes defined by mRS≥3.

The other primary outcome was in-hospital deterioration. The coding for deterioration included clinical condition worsening due to brain herniation, hemorrhagic transformation, neurological deterioration defined by an increase of 4 points or more in the NIHSS score compared to the admission score, and clinical deterioration due to medical problems. When there were specific causes for increases in the NIHSS scores by 4 points or more, such as brain herniation or hemorrhagic transformation, the patients were coded for these reasons; otherwise, we coded them for neurological deterioration. If mortality or critical conditions occurred owing to medical complications, we assigned them the code of in-hospital deterioration due to medical problems.

Features in the Models

The following categories of features were included in the models: (1) demographic features: age, sex, smoking habit, alcohol consumption, height, weight, and BMI; (2) medical comorbidities: a history of previous stroke, ischemic heart disease, congestive heart failure, atrial fibrillation, diabetes mellitus (DM), hypertension, and hyperlipidemia; (3) stroke-related index: NIHSS total score and subscores at ER and ward admission and stroke onset-to-hospitalization interval; (4) initial physiological parameters at admission: initial systolic blood pressure (SBP) and diastolic blood pressure, heart rate, respiratory rate, and body temperature; (5) initial laboratory parameters of blood tests: hemogram including the white blood cell (WBC) count and its differential counts, red blood cell (RBC) count, hemoglobin, hematocrit and platelet counts, prothrombin time (PT), activated partial thromboplastin time,

cholesterol and triglyceride profile, aspartate aminotransferase, alanine transaminase, blood urea nitrogen, creatinine, glucose, glycosylated hemoglobin, C-reactive protein, erythrocyte sedimentation rate, and homocysteine; (6) data of urine tests, including urine total protein and glucose levels.

Data Visualization

Unsupervised clustering provided an explicit grouping of the data, and direct visualization of the clusters showed the natural distribution of data. The t-distributed stochastic neighbor embedding (t-SNE) is a nonlinear dimensionality reduction for visualization [13]. Let P be the joint probability distribution for high dimension, and Q for low dimension. The distance between the 2 similarity matrices could be expressed as:

A gradient descent was performed to minimize this score, and the gradient could be computed as:

Machine Learning Models

Support Vector Machine (SVM)

The SVM was used to construct a hyperplane to split the data into 2 classes and optimize the distance between all data points and the hyperplane [14]. For a set of $\{x_i, y_i\}$, $i = 1, \ldots, N$, $x_i \in Rd y_i \in \{+1, -1\}$, the SVM found a vector ω such that $y_i (\omega T x_i - b) > 0$. The vector split the data into 2 classes. Many lines were available for splitting the set. The SVM optimized the solution by solving:

And retrieved the solution from:

Random Forest (RF)

The RF algorithm was based on bagging and decision [15]. Bootstrap aggregating (bagging) used repeated random sampling and replaced the training set to create a subset, reduce variance, and improve accuracy. Each subset of the training set conducted a random selection with features. The aggregation combined all predictions and yielded the regression mean and classification mode.

Light Gradient Boosting Machine (LGBM)

LGBM is a gradient boosting framework using tree-based learning algorithms [16]. In LGBM, gradient-based one-side sampling (GOSS) and exclusive feature bundling (EFB) were the 2 main techniques to improve efficiency and scalability. GOSS kept those data with large gradients and randomly dropped those with small gradients and reduced the calculation cost. EFB bundled exclusive features to reduce feature dimensions. The feature bundles could improve training efficiency without losing accuracy.

Deep Neural Network (DNN)

The DNN model was trained with tuned parameters in neurons by adjusting their weights and bias values to make the model's output closer to the ground truth [17]. If we set θ as all the parameters of the model and the input as x passing the neural network, $F(\theta)$, the output layer would generate the corresponding $F(x, \theta) = \hat{y}$. The embedding layer turned positive integers (indexes) into fixed-size vectors. The technique could avoid the sparse matrix obtained during the transformation of high-dimensional data into lower-dimensional data and turned categorical data into one-hot encoding data. In the DNN, the gradient descent algorithm solved the optimization problem by calculating the gradient of the loss function, updating the model's parameters in the opposite direction, and minimizing the loss. By selecting an optimal learning rate, a local minimum would be reached by iterations. Additional methods to optimize the model included batch normalization, which normalized the means and variances of each layer's inputs [18]. Dropout avoided overfitting by randomly omitting a certain fraction of neurons on each training case [19].

Data Processing

We applied a min-max normalizer to the numerical data for data engineering, split the data into 5 folds, and performed 5-fold cross-validation for performance evaluation. Cross-validation is a suitable approach to estimate the performance of a model when the data set is small. During the process of 5-fold cross-validation, the data set was first divided into 5 groups; then, each group was used as an unseen testing set in turn, whereas the remainder of the data set served as the training set (Figure 1). Notably, for DNN and LGBM, 10% of the training set was used as the validation set (tuning set) to prevent the overfitting problem and ensure that the model is trained well. Finally, the mean and SD of the testing accuracy in 5 rounds were evaluated as performance metrics.

Figure 1. Data enrollment. After the initial enrollment of 3589 patients, data cleaning excluded 809 patients and left 2780 eligible patients. The enrolled data set underwent k-fold cross-validation. In 5 folds, the data set was randomly divided such that 80% was for training and 20% for testing in each fold. The results of cross-validation underwent performance comparison with the ground truth and are expressed as the area under the curve of the receiver operating characteristic curve. mRS: modified Rankin Scale; NIHSS: National Institutes of Health Stroke Scale.
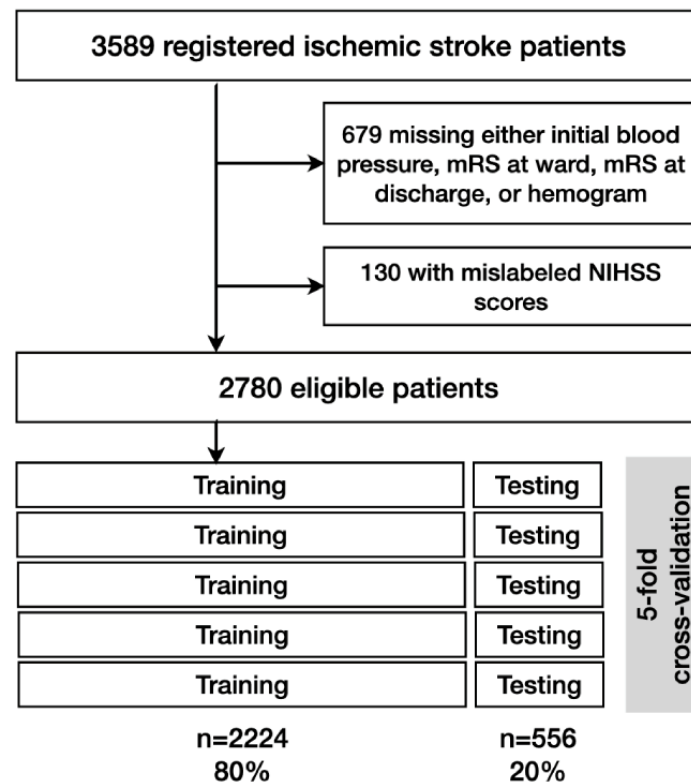
Transformation of the multiclassification model to a binary model was performed to improve model performance. After training the RF and LGBM multiclassifiers, the output summed up the mRS outcome {0,1,2} as False, and mRS outcome {3,4,5,6} as True.

Between-model comparison was conducted to rate the SVM, RF, LGBM, and DNN. Model performance in terms of the prediction ability was evaluated using the average area under the curve (AUC) of the receiver operating characteristic (ROC) curve; clear interpretations of true positives and false positives were essential for the classification problem.

Imbalanced Data

To handle imbalanced outcomes, 3 resampling methods were applied to make the 2 outcome classes more balanced. First, random under sampling (RUS) randomly dropped data from the majority class and often led to missing critical data. Second, random over sampling (ROS) randomly duplicated data of the minority class but sometimes led to overfitting of the minor samples. The third resampling method was the synthetic minority over-sampling technique (SMOTE) [20], which synthesized data from the minority class. The synthetic sample x is a point along the line segment joining x_i and x^i, where $\boldsymbol{x}^0_i = \boldsymbol{x}_i + (\boldsymbol{\hat{x}}_i - \boldsymbol{x}_i) \times \delta$ and the random number $\delta \in (0,1)$. The synthetic minority over-sampling technique-nominal continuous (SMOTE-NC) technique is the advanced modification of SMOTE and capable of handling mixed data sets of continuous and nominal features. The SMOTE-NC ran median

computations for nominal features and nearest neighbor computations for mixed data. The algorithm gave those nominal features the value occurring in most k-nearest neighbors.

Interpretation of Models

The SHAP, inspired by the Shapley value in game theory, assigned each feature a value of importance for a particular prediction [21]. The SHAP summary used kernel SHAP to estimate the Shapley value and visualized the prediction distribution among the feature values. For example, when approximating the original model f for a specific input x, local accuracy required the explanation model to match the output off for the simplified input x' that corresponded to the original input x:

$$\times$$

Data Availability

Anonymized data not published within this article will be made available on request from any qualified investigator under the regulations of our institutional review board.

Results

Data Enrollment

Initial screening identified 3589 patients of admission due to acute ischemic stroke. The data cleaning steps excluded 679 patients for missing records of blood pressure, mRS, and hemograms. Another 130 patients were excluded for mislabeled

NIHSS scores. The missing rate of all the features was under 10%. A total of 2780 eligible patients were enrolled. The data underwent 5-fold cross-validation. In each fold, the models randomly divided the whole data set into 80% data for training and 20% data for testing. The performance in each fold was compared with the ground truth and quantified in the AUC of ROC curves. The final AUC results were the means and SDs obtained from the 5-fold cross-validation (Figure 1).

Prediction of mRS at Hospital Discharge

The t-SNE was used for unsupervised clustering to visualize the data. Of the entire data set containing 2780 cases, the 1284 orange dots for a bad outcome and the 1571 blue dots for a good outcome overlapped to a certain degree (Figure 2A). The t-SNE results showed the relationship between the bad and good outcomes at the feature stage, but this does not mean that the machine learning models could not separate the mixed data.

Figure 2. Prediction of modified Rankin Scale (mRS) at hospital discharge. The outcome variable mRS at discharge was transformed from 6 ordinal classes to a binary class. The good outcome was defined by mRS {0,1,2}, whereas the bad outcome was indicated by mRS {3,4,5,6}. (A) The t-SNE graph shows the distribution of the data. Orange indicates discharge mRS 3-6 and blue represents mRS 0-2. (B) ROC curves for 4 machine learning models. (C) Comparisons of AUC between the data with and without normalization of numerical features. (D) AUC for different amounts of data. AUC: area under the curve; DNN: deep neural network; LGBM; light gradient boosting machine; mRS: modified Rankin Scale; RF: random forest; ROC: receiver operating characteristic; SVM: support vector machine; t-SNE: t-distributed stochastic neighbor embedding.

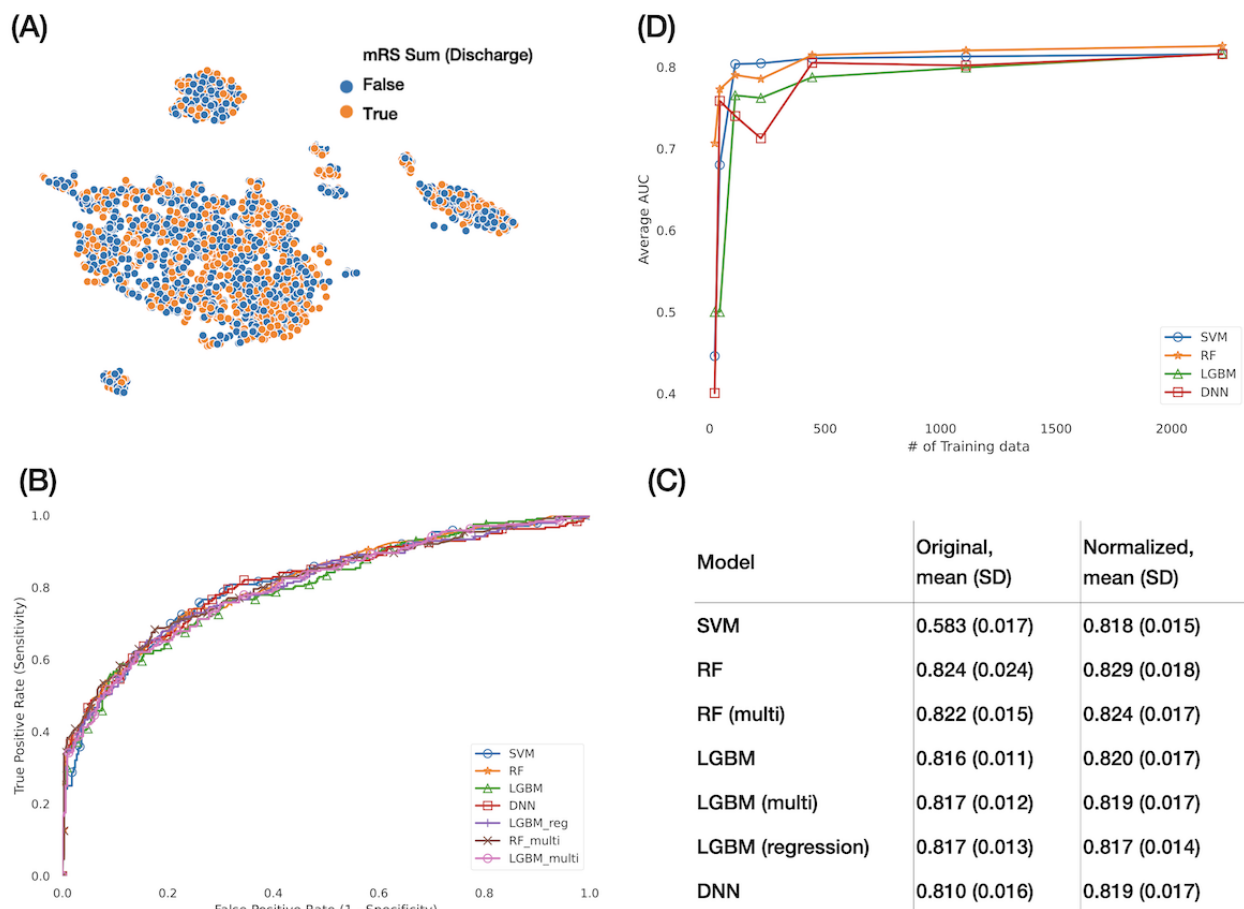| Model	Original, mean (SD)	Normalized, mean (SD)
SVM	0.583 (0.017)	0.818 (0.015)
RF	0.824 (0.024)	0.829 (0.018)
RF (multi)	0.822 (0.015)	0.824 (0.017)
LGBM	0.816 (0.011)	0.820 (0.017)
LGBM (multi)	0.817 (0.012)	0.819 (0.017)
LGBM (regression)	0.817 (0.013)	0.817 (0.014)
DNN	0.810 (0.016)	0.819 (0.017)

Figure 2B shows the ROC curve for comparing model performances using normalized data. The overlapping curves indicate that the models performed equally well with the AUC being approximately 0.8, with no model being significantly superior to the others. Normalization of the numerical data improved the performance of the SVM model because of its linear nature, but normalization was not beneficial for the tree models and DNN (Figure 2C). We further simulated different volumes of data by sampling different fractions (0.01, 0.02, 0.05, 0.1, 0.2, and 0.5) of data from the entire training data set, conducted the 5-fold cross-validation, and determined the performance at each data volume (Figure 2D). On increasing the training data to more than 500 samples, the model performance reached a plateau, with the average AUC for RF

being near 0.8, almost as high as that for the entire data. With more data, the performance of all the 4 models improved. In contrast, with limited data, the performance would also be acceptable.

We further applied feature importance and compared it with SHAP in terms of the summary aspect. The top 5 features in RF and LGBM were similar in terms of the NIHSS total score, age, WBC differential counts of lymphocyte and segmented neutrophil, and renal function creatinine (Figure 3A). On the other hand, the SHAP summary of the RF and LGBM models presented the ranking of important features and their influence on predicting outcomes (Figures 3B-3C). For example, the SHAP summary suggested higher NIHSS total scores, worse lower limb motor function, older age, higher segmented

neutrophil, and lower lymphocyte percentage of WBC differential counts, indicating a higher mRS score for more

dependency at hospital discharge.

Figure 3. Feature importance for predicting modified Rankin Scale at hospital discharge. (A) Top 5 important features of random forest and light gradient boosting machine. SHapley Additive exPlanations of (B) random forest and (C) light gradient boosting machine. Red indicates higher feature sample values, and blue indicates lower feature sample values. For example, the higher the total National Institutes of Health Stroke Scale scores at emergency room and at ward admission, the more severe would be the stroke outcome. ALT: alanine transaminase; APTT: activated partial thromboplastin time; DM: diabetes mellitus; ER: emergency room; LGBM: light gradient boosting machine; LOC: level of consciousness; NIHSS: National Institutes of Health Stroke Scale; RF: random forest; SHAP: SHapley Additive exPlanations. Wd: ward.
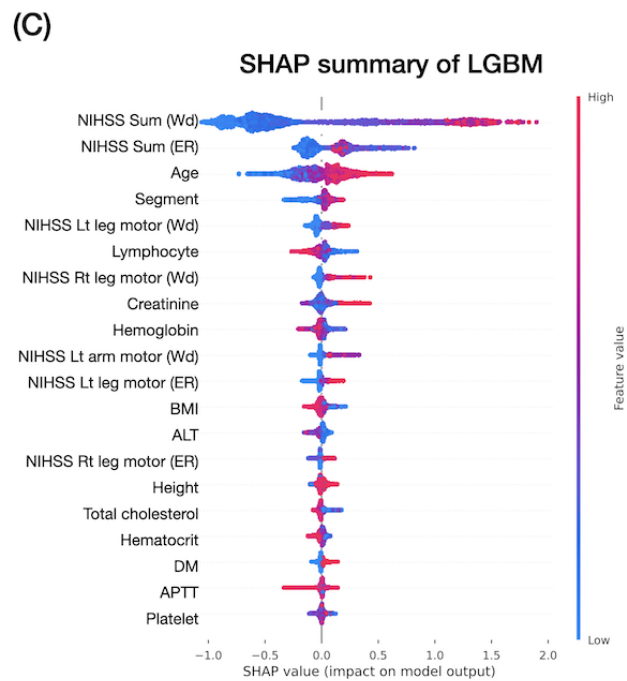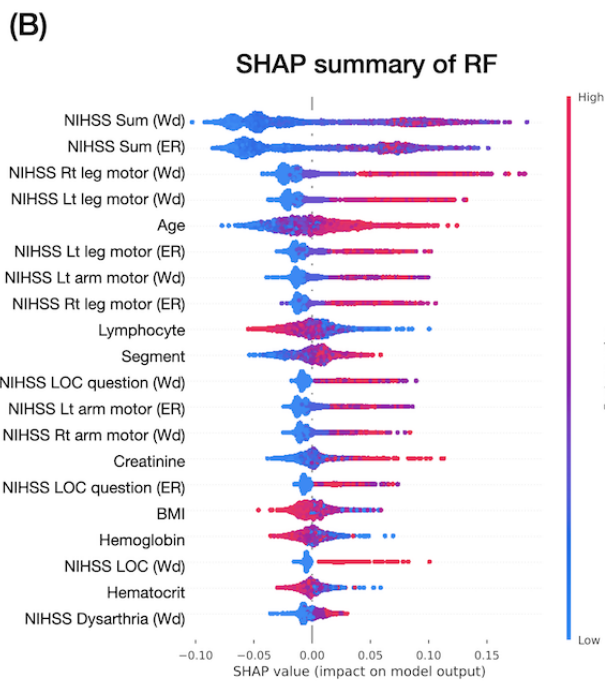

Prediction of In-Hospital Deterioration

Of the initial cohort of 2780 patients, 2622 (94%) were nondeterioration and 158 (6%) were deterioration cases. The coding ratio of in-hospital neurological deterioration, medical problems, brain herniation, and hemorrhagic transformation was 0.64:0.18:0.14:0.04. Next, we compared the performances of the 4 models in predicting deterioration and the 4 resampling

methods for imbalanced data. Finally, we compared the feature importance.

The sample grouped and visualized by t-SNE showed that deteriorations were the minority surrounded by nondeterioration samples (Figure 4A). The resampling methods RUS and ROS did not group samples well (Figures 4B-4C). Finally, the SMOTE-NC produced synthetic data in the neighborhood of true data, but the data were still not grouped well (Figure 4D).

Figure 4. Prediction of in-hospital deterioration. (A) Visualization by t-distributed stochastic neighbor embedding of the original sample shows an imbalanced outcome. The 3 resampling methods processed the imbalanced data with (B) random under sampling decreasing the majority class, (C) random over sampling increasing the minority class, and (D) synthetic minority over-sampling technique with nominal continuous data synthesis from the minority class. (E) Receiver operating characteristic curves for predicting in-hospital deterioration from the data without resampling. (F) Comparison of the area under the curve in the different resampling methods. Random under sampling was a reasonable choice for resampling. It improved the performance of the random forest, light gradient boosting machine, and support vector machine models, but not the deep neural network. The deep neural network performed better on the original data set than on the resampled data set. DNN: deep neural network; LGBM: light gradient boosting machine; RF: random forest; ROC: receiver operating characteristic; ROS: random over sampling; RUS: random under sampling; SMOTE-NC: synthetic minority over-sampling technique-nominal continuous; SVM: support vector machine.
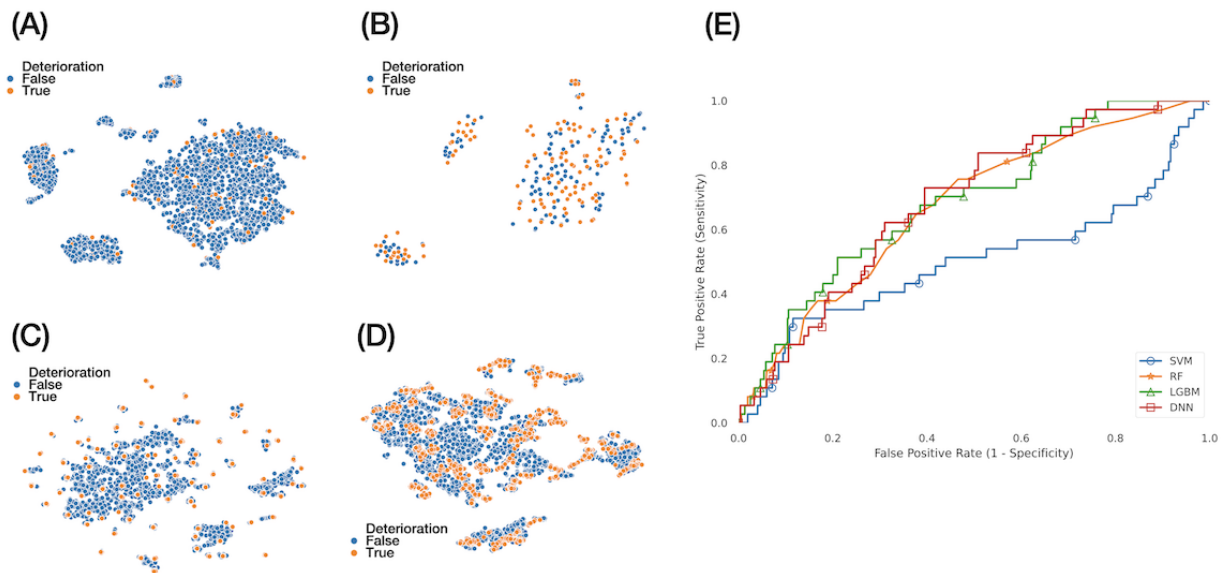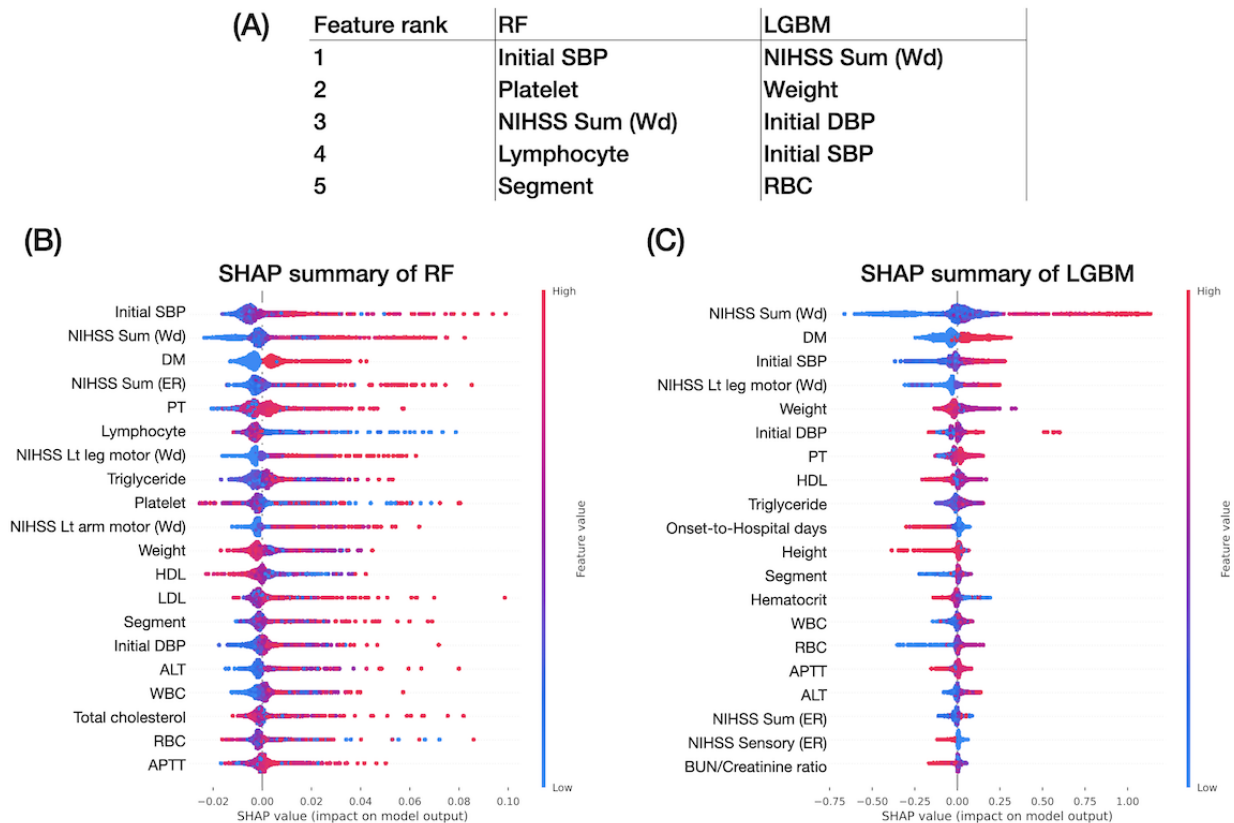

(F)

Model	No resampling, mean (SD)	RUS, mean (SD)	ROS, mean (SD)	SMOTE-NC, mean (SD)
RF	0.710 (0.023)	0.738 (0.036)	0.666 (0.019)	0.589 (0.065)
LGBM	0.683 (0.047)	0.701 (0.068)	0.619 (0.028)	0.590 (0.061)
SVM	0.483 (0.041)	0.710 (0.060)	0.663 (0.027)	0.509 (0.087)
DNN	0.732 (0.064)	0.688 (0.035)	0.511 (0.079)	0.624 (0.036)

The ROC curves showed the predictive performance for in-hospital deterioration of different models. In the original data set, RF and DNN outperformed SVM and LGBM (Figure 4E, data without resampling). As for each resampling method, RUS improved the performance of all the models except DNN (Figure 4F). The DNN model performed better on the original data set than on resampled data. The performance of SVM was significantly improved by RUS, ROS, and SMOTE-NC.

We further compared the top 5 important features with nonresampling data (Figure 5A). The NIHSS total score was

critical for predicting in-hospital deterioration. In the SHAP summary, we learned that the higher the NIHSS score, the higher the risk of deterioration. Notably, the initial SBP was prominent in the top 5 important features of RF and LGBM (Figure 5A) and their SHAP summaries (Figures 5B-5C). The SHAP summaries of RF and LGBM showed that the higher the initial SBP, the higher the risk of in-hospital deterioration. In addition, the features obtained from the blood test hemograms, including WBC differential count, platelet count, PT, and RBC, appeared in the top features. Having DM was also crucial in predisposing in-hospital deterioration (Figures 5B-5C).

Figure 5. Feature importance for predicting in-hospital deterioration (without resampling). (A) Top 5 important features include initial systolic blood pressure at hospital admission in random forest and light gradient boosting machine. National Institutes of Health Stroke Scale total score at ward admission is also an important feature in both models. SHapley Additive exPlanations of (B) random forest and of (C) light gradient boosting machine. ALT: alanine transaminase; APTT: activated partial thromboplastin time; BUN: blood urea nitrogen; DBP: diastolic blood pressure; DM: diabetes mellitus; ER: emergency room; HDL: high-density lipoprotein; LDL: low-density lipoprotein; NIHSS: National Institutes of Health Stroke Scale; PT: prothrombin time; RBC: red blood cell; SBP: systolic blood pressure; WBC: white blood cell; Wd: ward.

(A)

Feature rank	RF	LGBM
1	Initial SBP	NIHSS Sum (Wd)
2	Platelet	Weight
3	NIHSS Sum (Wd)	Initial DBP
4	Lymphocyte	Initial SBP
5	Segment	RBC

Discussion

Summary

In this study, we used machine learning to predict the mRS outcome at hospital discharge and in-hospital deterioration in the setting of acute ischemic stroke. RF performed the best in most tasks. Applying SHAP to the models combining numerical and higher-dimensional features was feasible, and the SHAP summary emphasized the importance of these features for clinical explanations. As for the resampling of imbalanced data, the effects of resampling on the performance improvement of the models were only equivocal, and SMOTE-NC was not an outstanding method.

Several studies compared models for stroke outcome prediction. In a study with data of over 15,000 patients, DNN outperformed traditional methods when predicting stroke patient mortality [22]. The stroke outcomes predicted by DNN were superior to the ASTRAL scores [23]. However, DNN made no difference in another study predicting 3-month mRS [23]. In our study, DNN did not excel in predicting the discharge mRS, but it performed better than the other models when predicting in-hospital deterioration using nonresampling data. Therefore, DNN is a reasonable choice for the prediction of early deterioration in acute ischemic stroke.

Gradient boosting machine (GBM) and RF are tree-based machine learning models. In a comparative study, extreme gradient boosting (XGBoost) performed better than the traditional GBM in predicting 3-month mRS [24]. However, another study has mentioned that RF performs the best when compared to XGBoost and other traditional models, such as logistic regression, decision tree, and SVM [25]. Similarly, we found that RF performed well in targeting in-hospital deterioration and predicting independence at discharge. RF is effective with imbalanced data and therefore performs well in medical issues with scarce outcomes [26]. RF is suitable for predicting medical diagnosis, and feature ranking helps the RF model in medical classification [27]. Therefore, using RF in predicting early stroke outcomes was feasible. On the contrary, SVM is the least suitable model for stroke early outcome prediction.

Recent Progress in Model Interpretation

Interpreting how models predict outcomes is sometimes as crucial as their accuracy. In recent years, there has been an increasing amount of literature explaining machine learning models, which helps investigate their learning mechanisms, debug these models, avoid adversarial attacks, and verify the fairness and bias of these models [28,29]. Tree models have some simple inbuilt methods, such as counts for the features used in the model. However, these methods lead to biased

approaches, as they tend to inflate the importance of continuous features or high-cardinality categorical variables. To solve the black-box nature of complex models such as deep learning models, the additive feature attribution methods alter the inputs to see how the outputs react and provide a practical solution for the models [30]. The local interpretable model-agnostic explanation, introduced in 2016, approximates a black-box model using a simple linear surrogate model locally [31,32]. Recent explainers, including the SHAP announced in 2018, explore the model from a more global perspective. [21,32]. In a study aiming to predict extubating failure in intensive care units, SHAP analysis proved effective and accurate [33]. With the help of SHAP, we determined the contribution of each feature toward predicting stroke outcomes. The SHAP summary distinguished the features that could separate targets and nontargets from those features that could not.

When working with imbalanced data, SMOTE resampling often achieves better performance in predicting stroke occurrence [34]. However, investigating important features with synthetic data maybe not be persuasive because of its nature of linear interpolation. Repeatedly resampling categorical features could lead to overfitting of the synthetic data. In contrast, continuous features usually stood out without resampling. SMOTE-NC resampling for the imbalanced data of in-hospital deterioration could even worsen model performance. The reason may be the overfitting of categorical data (Figure 4F).

Initial Blood Pressure in Predicting Early Outcomes of Ischemic Stroke

This work followed the SRICHS registry study, which found the associations between initial blood pressure and 1-year outcomes [35]. In this work, the machine learning models RF and LGBM identified high initial SBP as a crucial factor influencing in-hospital deterioration. High SBP is a strong predictor of stroke [36] and ranks the first among the stroke risk factors contributing to stroke-related DALYs [37]. Chronic hypertension is the most important modifiable risk factor of stroke, according to the INTERSTROKE study [38]. Persistent high blood pressure indicates a worse long-term stroke outcome [39]. High initial blood pressure is detrimental to early neurological outcomes and heralds the deterioration of neurological function in the hospital [40]. Patients with high blood pressure tended to encounter acute infarct volume expansion [41]. Consistent with traditional statistics, our machine learning models supported the importance of blood pressure in predicting early deteriorations in terms of neurological, pathophysiological, and medical changes of acute ischemic stroke. During the creation of this data set, endovascular therapy was not a standard treatment yet. Current studies highlight the importance of blood pressure for stroke patients receiving endovascular therapy [42]. Possessing the capability to process complex data, our machine learning models are promising tools to solve complicated problems in the new era of stroke care, such as blood pressure problems in endovascular therapy.

DM and Early Stroke Outcomes

DM is a known risk factor for stroke. It accelerates the development of ischemic stroke at a younger age [43].

Compared to nondiabetic stroke patients, ischemic stroke patients with DM had worse neurological deficits, less favorable outcomes from rehabilitation, delayed recovery from the stroke-related deficit, a longer hospital stay for acute ischemic stroke, a higher probability of experiencing a recurrent stroke within 1 year, and a higher rate of 1-year mortality [43,44]. In our study, having DM was a strong predictor for in-hospital deterioration in the SHAP summary of RF and LGBM. Other studies also revealed that DM predisposed early neurological deterioration [45] and increased mortality during hospital stay [43]. This finding suggests that the explainable machine learning model using the SHAP summary is as informative as the stroke registry statistics.

Limitations of the Study

There were several limitations of this study. First, the registry-based study might have inconsistent assessments and treatments of the patients, incomplete data registration, missing outcomes, and loss of follow-up data [46]. Because of the potentially underreported data, the outcomes might be underestimated. Still, tracking the natural history of a disease, collecting a large number of patients, and yielding generalizable findings make registry-based studies valuable in understanding diseases and outcome assessments. Second, our machine learning models predicted discharge mRS more accurately than in-hospital deterioration. Because general condition deterioration involves multiple factors and individual circumstances, predicting it is more complicated than predicting the neurological status at discharge, which could refer to the initial neurological status. The attributes of the current study design limited the quality and quantity of the features used in model design. In future studies, prospectively collecting delicate parameters, such as continuous vital sign recordings and neuroimages, may improve the performance of these models when predicting in-hospital deterioration. Third, the data set we used in this study was collected in 2009. In the past 10 years, the disease course of ischemic stroke may have changed due to the popularity of comorbidities, demography of stroke proneness, progress in stroke treatment, and improved poststroke care. The machine learning models used in this study may not be completely suitable for new data, and the models may need to be retrained and adjusted. Nevertheless, novel therapies, such as intravenous thrombolysis and endovascular thrombectomy, for acute ischemic stroke were not prevalent a decade ago, and, therefore, we could clearly understand the disease nature course from this data analysis.

Conclusions

RF, an ensemble algorithm of regression and classification containing multiple decision trees, outperformed SVM, LGBM, and DNN in targeting early stroke outcomes of discharge mRS. RF and DNN performed well in predicting in-hospital deterioration. Using the SHAP summary and feature importance ranking may help clinicians in explaining the prediction of the machine learning models. The multidomain feature bank, combining physiological monitoring values, laboratory data, and neurological severities, as well as the improved performance of the models helped predict in-hospital deterioration. These

machine learning models are promising for advanced applications in stroke outcome prediction.

Conflicts of Interest

None declared.

References

1. Murray CJ, Lopez AD. Measuring the global burden of disease. N Engl J Med 2013 Aug;369(5):448-457. [doi: 10.1056/nejmra1201534]
2. Goyal M, Ospel JM, Kappelhof M, Ganesh A. Challenges of outcome prediction for acute stroke treatment decisions. Stroke 2021 May;52(5):1921-1928. [doi: 10.1161/strokeaha.120.033785]
3. Powers W, Rabinstein A, Ackerson T, Adeoye OM, Bambakidis N, Becker K, American Heart Association Stroke Council. 2018 guidelines for the early management of patients with acute ischemic stroke: a guideline for healthcare professionals from the American Heart Association/American Stroke Association. Stroke 2018 Mar;49(3):e46-e110 [FREE Full text] [doi: 10.1161/STR.0000000000000158] [Medline: 29367334]
4. Adams HP, Davis PH, Leira EC, Chang K, Bendixen BH, Clarke WR, et al. Baseline NIH Stroke Scale score strongly predicts outcome after stroke: a report of the Trial of Org 10172 in Acute Stroke Treatment (TOAST). Neurology 1999 Jul;53(1):126. [doi: 10.1212/wnl.53.1.126] [Medline: 10408548]
5. Reid JM, Gubitz GJ, Dai D, Kydd D, Eskes G, Reidy Y, et al. Predicting functional outcome after stroke by modelling baseline clinical and CT variables. Age Ageing 2010 May;39(3):360-366. [doi: 10.1093/ageing/afq027] [Medline: 20233732]
6. Ntaios G, Gioulekas F, Papavasileiou V, Strbian D, Michel P. ASTRAL, DRAGON and SEDAN scores predict stroke outcome more accurately than physicians. Eur J Neurol 2016 Jul;23(11):1651-1657. [doi: 10.1111/ene.13100]
7. Tran BX, Latkin CA, Vu GT, Nguyen HLT, Nghiem S, Tan M, et al. The current research landscape of the application of artificial intelligence in managing cerebrovascular and heart diseases: a bibliometric and content analysis. Int J Environ Res Public Health 2019 Jul;16(15):2699 [FREE Full text] [doi: 10.3390/ijerph16152699] [Medline: 31362340]
8. Winzeck S, Hakim A, McKinley R, Pinto JAADSR, Alves V, Silva C, et al. ISLES 2016 and 2017-benchmarking ischemic stroke lesion outcome prediction based on multispectral MRI. Front Neurol 2018 Sep;9:679 [FREE Full text] [doi: 10.3389/fneur.2018.00679] [Medline: 30271370]
9. Bacchi S, Oakden-Rayner L, Zerner T, Kleinig T, Patel S, Jannes J. Deep learning natural language processing successfully predicts the cerebrovascular cause of transient ischemic attack-like presentations. Stroke 2019 Mar;50(3):758-760. [doi: 10.1161/strokeaha.118.024124]
10. Arts DL, Abu-Hanna A, Medlock SK, van Weert HCPM. Effectiveness and usage of a decision support system to improve stroke prevention in general practice: a cluster randomized controlled trial. PLoS One 2017 Feb;12(2):e0170974 [FREE Full text] [doi: 10.1371/journal.pone.0170974] [Medline: 28245247]
11. Lee T, Chang C, Chang Y, Chang K, Chung J. Establishment of electronic chart-based stroke registry system in a medical system in Taiwan. J Formos Med Assoc 2011 Aug;110(8):543-547. [doi: 10.1016/s0929-6646(11)60081-8]
12. Sulter G, Steen C, Jacques De Keyser. Use of the Barthel index and modified Rankin Scale in acute stroke trials. Stroke 1999 Aug;30(8):1538-1541. [doi: 10.1161/01.str.30.8.1538]
13. Maaten L, Hinton G. Visualizing data using t-sne. J Mach Learn Res 2008:2579-2605.
14. Lauer F, Bloch G. Incorporating prior knowledge in support vector machines for classification: a review. Neurocomputing 2008 Mar;71(7-9):1578-1594. [doi: 10.1016/j.neucom.2007.04.010]
15. Breiman L. Random forests. Mach Learn 2001 Oct;45:5-32. [doi: 10.1007/978-1-4899-7687-1_695]
16. Ke G, Meng Q, Finley T, Wang T, Chen W, Ma W. Lightgbm: a highly efficient gradient boosting decision tree. In: Advances in Neural Information Processing Systems 30. 2017 Presented at: 31st International Conference on Neural Information Processing Systems; 2017 December 4-9, 2017; Long Beach, United States p. 3149-3157.
17. Liu W, Wang Z, Liu X, Zeng N, Liu Y, Alsaadi FE. A survey of deep neural network architectures and their applications. Neurocomputing 2017 Apr;234:11-26. [doi: 10.1016/j.neucom.2016.12.038]
18. Ioffe S, Szegedy C. Batch normalization: accelerating deep network training by reducing internal covariate shift. In: Proceedings of the 32nd International Conference on Machine Learning. 2015 Presented at: 32nd International Conference on Machine Learning; July 6-11, 2015; Lille, France p. 448-456.

19. Srivastava N, Hinton G, Krizhevsky A, Sutskever I, Salakhutdinov R. Dropout: a simple way to prevent neural networks from overfitting. J Mach Learn Res 2014;15:1929-1958.

20. Chawla NV, Bowyer KW, Hall LO, Kegelmeyer WP. SMOTE: synthetic minority over-sampling technique. J Artif Intell Res 2002 Jun;16:321-357. [doi: 10.1613/jair.953]

21. Lundberg SM, Lee SI. A unified approach to interpreting model predictions. In: Advances in Neural Information Processing Systems 30. 2017 Presented at: 31st International Conference on Neural Information Processing Systems; December 4-9, 2017; Long Beach, United States p. 4768-4777.

22. Cheon S, Kim J, Lim J. The use of deep learning to predict stroke patient mortality. Int J Environ Res Public Health 2019 May;16(11):1876 [FREE Full text] [doi: 10.3390/ijerph16111876] [Medline: 31141892]

23. Heo J, Yoon JG, Park H, Kim YD, Nam HS, Heo JH. Machine learning–based model for prediction of outcomes in acute stroke. Stroke 2019 May;50(5):1263-1265. [doi: 10.1161/strokeaha.118.024293]

24. Xie Y, Jiang B, Gong E, Li Y, Zhu G, Michel P, et al. Use of gradient boosting machine learning to predict patient outcome in acute ischemic stroke on the basis of imaging, demographic, and clinical information. AJR Am J Roentgenol 2019 Jan;212(1):44-51. [doi: 10.2214/AJR.18.20260]

25. Monteiro M, Fonseca AC, Freitas AT, Pinho e Melo T, Francisco AP, Ferro JM, et al. Using machine learning to improve the prediction of functional outcome in ischemic stroke patients. IEEE/ACM Trans Comput Biol Bioinf 2018 Nov;15(6):1953-1959. [doi: 10.1109/TCBB.2018.2811471]

26. Khalilia M, Chakraborty S, Popescu M. Predicting disease risks from highly imbalanced data using random forest. BMC Med Inform Decis Mak 2011 Jul;11(1):51 [FREE Full text] [doi: 10.1186/1472-6947-11-51] [Medline: 21801360]

27. Alam MZ, Rahman MS, Rahman MS. A random forest based predictor for medical data classification using feature ranking. Inform Med Unlocked 2019;15:100180. [doi: 10.1016/j.imu.2019.100180]

28. Montavon G, Samek W, Müller K. Methods for interpreting and understanding deep neural networks. Digit Signal Process 2018 Feb;73:1-15. [doi: 10.1016/j.dsp.2017.10.011]

29. Samek W, Montavon G, Vedaldi A, Hansen L. Explainable AI: Interpreting, Explaining and Visualizing Deep Learning. Cham, Switzerland: Springer; 2019.

30. Gilpin L, Bau D, Yuan B, Bajwa A, Specter M, Kagal L. Explaining explanations: an overview of interpretability of machine learning. 2018 Presented at: IEEE 5th International Conference on data science and advanced analytics (DSAA); October 1-4, 2018; Turin, Italy p. 80-89. [doi: 10.1109/DSAA.2018.00018]

31. Ribeiro M, Singh S, Guestrin C. "Why Should I Trust You?" Explaining the predictions of any classifier. 2016 Aug Presented at: 22nd ACM SIGKDD international Conference on Knowledge Discovery and Data Mining; August 13-17, 2016; San Francisco, United States p. 1135-1144. [doi: 10.1145/2939672.2939778]

32. Stiglic G, Kocbek P, Fijacko N, Zitnik M, Verbert K, Cilar L. Interpretability of machine learning‐based prediction models in healthcare. WIREs Data Mining Knowl Discov 2020 Sep;10(5):e1379. [doi: 10.1002/widm.1379]

33. Chen T, Xu J, Ying H, Chen X, Feng R, Fang X, et al. Prediction of extubation failure for intensive care unit patients using light gradient boosting machine. IEEE Access 2019 Oct;7:150960-150968. [doi: 10.1109/ACCESS.2019.2946980]

34. Wu Y, Fang Y. Stroke prediction with machine learning methods among older Chinese. Int J Environ Res Public Health 2020 Mar;17(6):1828 [FREE Full text] [doi: 10.3390/ijerph17061828] [Medline: 32178250]

35. Liu C, Wei Y, Lin J, Chang C, Chang T, Huang K, Stroke Registry in Chang Gung Healthcare System (SRICHS) Investigators. Initial blood pressure is associated with stroke severity and is predictive of admission cost and one-year outcome in different stroke subtypes: a SRICHS registry study. BMC Neurol 2016 Feb;16(1):27 [FREE Full text] [doi: 10.1186/s12883-016-0546-y] [Medline: 26923538]

36. Lindenstrøm E, Boysen G, Nyboe J. Influence of systolic and diastolic blood pressure on stroke risk: a prospective observational study. Am J Epidemiol 1995 Dec;142(12):1279-1290. [doi: 10.1093/oxfordjournals.aje.a117595] [Medline: 7503048]

37. Feigin VL, Roth GA, Naghavi M, Parmar P, Krishnamurthi R, Chugh S, et al. Global burden of stroke and risk factors in 188 countries, during 1990–2013: a systematic analysis for the Global Burden of Disease Study 2013. The Lancet Neurology 2016 Aug;15(9):913-924. [doi: 10.1016/s1474-4422(16)30073-4]

38. O'Donnell MJ, Chin SL, Rangarajan S, Xavier D, Liu L, Zhang H, et al. Global and regional effects of potentially modifiable risk factors associated with acute stroke in 32 countries (INTERSTROKE): a case-control study. The Lancet 2016 Aug;388(10046):761-775. [doi: 10.1016/s0140-6736(16)30506-2]

39. Zheng X, Peng Y, Zhong C, Xie X, Wang A, Zhu Z, et al. Systolic blood pressure trajectories after discharge and long-term clinical outcomes of ischemic stroke. Hypertension 2021 May;77(5):1694-1702. [doi: 10.1161/hypertensionaha.120.16881]

40. Ishitsuka K, Kamouchi M, Hata J, Fukuda K, Matsuo R, Kuroda J, et al. High blood pressure after acute ischemic stroke is associated with poor clinical outcomes. Hypertension 2014 Jan;63(1):54-60. [doi: 10.1161/hypertensionaha.113.02189]

41. Castillo J, Leira R, Garcı́a MM, Serena J, Blanco M, Dávalos A. Blood pressure decrease during the acute phase of ischemic stroke is associated with brain injury and poor stroke outcome. Stroke 2004 Feb;35(2):520-526. [doi: 10.1161/01.str.0000109769.22917.b0]

42. Rasmussen M, Schönenberger S, Hendèn PL, Valentin JB, Espelund US, Sørensen LH, SAGA collaborators. Blood pressure thresholds and neurologic outcomes after endovascular therapy for acute ischemic stroke: an analysis of individual patient

data from 3 randomized clinical trials. JAMA Neurol 2020 May;77(5):622-631 [FREE Full text] [doi: 10.1001/jamaneurol.2019.4838] [Medline: 31985746]

43. Jørgensen H, Nakayama H, Raaschou HO, Olsen TS. Stroke in patients with diabetes. The Copenhagen Stroke Study. Stroke 1994 Oct;25(10):1977-1984. [doi: 10.1161/01.STR.25.10.1977] [Medline: 8091441]

44. Lau L, Lew J, Borschmann K, Thijs V, Ekinci EI. Prevalence of diabetes and its effects on stroke outcomes: a meta-analysis and literature review. J Diabetes Investig 2019 May;10(3):780-792 [FREE Full text] [doi: 10.1111/jdi.12932] [Medline: 30220102]

45. Tanaka R, Ueno Y, Miyamoto N, Yamashiro K, Tanaka Y, Shimura H, et al. Impact of diabetes and prediabetes on the short-term prognosis in patients with acute ischemic stroke. J Neurol Sci 2013 Sep;332(1-2):45-50. [doi: 10.1016/j.jns.2013.06.010]

46. Galluccio F, Walker UA, Nihtyanova S, Moinzadeh P, Hunzelmann N, Krieg T, et al. Registries in systemic sclerosis: a worldwide experience. Rheumatology (Oxford) 2011 Jan;50(1):60-68. [doi: 10.1093/rheumatology/keq355] [Medline: 21148153]

Abbreviations

ASTRAL: Acute Stroke Registry and Analysis of Lausanne
AUC: area under the curve
DALYs: disability-adjusted life years
DM: diabetes mellitus
DNN: deep neural network
EFB: exclusive feature bundling
ER: emergency room
GOSS: gradient-based one-side sampling
LGBM: light gradient boosting machine
mRS: modified Rankin Scale
NIHSS: National Institutes of Health Stroke Scale
PT: prothrombin time
RBC: red blood cell
RF: random forest
ROC: receiver operating characteristic
ROS: random over sampling
RUS: random under sampling
SBP: systolic blood pressure
SHAP: SHapley Additive exPlanations
SMOTE: synthetic minority over-sampling technique
SMOTE-NC: synthetic minority over-sampling technique-nominal continuous
SRICHS: Stroke Registry of the Chang Gung Healthcare System
SVM: support vector machine
t-SNE: t-distributed stochastic neighbor embedding
WBC: white blood cell

Original Paper

A Comparison of Different Modeling Techniques in Predicting Mortality With the Tilburg Frailty Indicator: Longitudinal Study

Tjeerd van der Ploeg[1], PhD; Robbert Gobbens[1,2,3], PhD

[1]Faculty of Health, Sports and Social Work, Inholland University of Applied Sciences, Amsterdam, Netherlands

[2]Zonnehuisgroep Amstelland, Amstelveen, Netherlands

[3]Department Family Medicine and Population Health, Faculty of Medicine and Health Sciences, University of Antwerp, Antwerp, Belgium

Corresponding Author:
Tjeerd van der Ploeg, PhD
Faculty of Health, Sports and Social Work
Inholland University of Applied Sciences
De Boelelaan 1109
Amsterdam, 1081 HV
Netherlands
Phone: 31 653519264
Email: tvdploeg@quicknet.nl

Related Article:

This is a corrected version. See correction statement: https://medinform.jmir.org/2022/4/e31479/

Abstract

Background: Modern modeling techniques may potentially provide more accurate predictions of dichotomous outcomes than classical techniques.

Objective: In this study, we aimed to examine the predictive performance of eight modeling techniques to predict mortality by frailty.

Methods: We performed a longitudinal study with a 7-year follow-up. The sample consisted of 479 Dutch community-dwelling people, aged 75 years and older. Frailty was assessed with the Tilburg Frailty Indicator (TFI), a self-report questionnaire. This questionnaire consists of eight physical, four psychological, and three social frailty components. The municipality of Roosendaal, a city in the Netherlands, provided the mortality dates. We compared modeling techniques, such as support vector machine (SVM), neural network (NN), random forest, and least absolute shrinkage and selection operator, as well as classical techniques, such as logistic regression, two Bayesian networks, and recursive partitioning (RP). The area under the receiver operating characteristic curve (AUROC) indicated the performance of the models. The models were validated using bootstrapping.

Results: We found that the NN model had the best validated performance (AUROC=0.812), followed by the SVM model (AUROC=0.705). The other models had validated AUROC values below 0.700. The RP model had the lowest validated AUROC (0.605). The NN model had the highest optimism (0.156). The predictor variable "difficulty in walking" was important for all models.

Conclusions: Because of the high optimism of the NN model, we prefer the SVM model for predicting mortality among community-dwelling older people using the TFI, with the addition of "gender" and "age" variables. External validation is a necessary step before applying the prediction models in a new setting.

KEYWORDS

modeling techniques; area under the receiver operating characteristic curve; bootstrapping; validation; predictor variable importance

XSL•FO
RenderX

Introduction

Predicting the survival probability of patients is important for various purposes in biomedaical research, such as patient counseling, medical decision-making, and benchmarking. The traditional analysis of survival problems uses Kaplan-Meier analysis and Cox regression modeling to predict the survival probability depending on various predictor variables.

Prediction is complicated by the specification of the model structure, such as the inclusion of main effects, potential nonlinearities, and statistical interaction [1-3]. While most prediction models for binary endpoints are still based on logistic regression (LR) analysis, there is increasing interest in other, more modern techniques, such as neural networks (NNs), random forests (RFs), and support vector machines (SVMs). These techniques hold the promise of better capturing nonlinearities and interactions in medical data and are, therefore, attractive in possibly providing better predictions [4].

NNs were used in 1998 for the analysis of survival data [5], and in 2007, applications of random survival forests were described [6]. SVMs were used in the context of breast cancer survival and chemotherapy [7]. In 2009, prognostic indexes were compared using modern techniques and Cox regression analysis in breast cancer data [8].

The aim of this study was to determine the best modeling technique for the prediction of mortality in a sample of community-dwelling older people by components of frailty using a follow-up period of 7 years. Frailty is the focus of much attention in practice, policy, and research. This is hardly surprising, since frailty in older people is predictive for disability [9], an increase in health care use [10], lower quality of life, and mortality [11].

Frailty is often operationalized by physical components, for example, in the phenotype of frailty by Fried et al [9]. However, only paying attention to physical limitations that older people may have or experience can lead to fragmentation of care [12] and then, potentially, to a reduction of quality of care and a decrease in quality of life of older people. Therefore, we used the Tilburg Frailty Indicator (TFI), a multidimensional scale including physical, psychological, and social components, for assessing frailty [13]. The TFI was developed on the basis of an extensive literature review and consultation with experts [12-14] and has shown good psychometric properties [15].

Five studies have examined the predictive value of the TFI for mortality [16-20]. Only one of these previous studies used the original TFI and conducted the study among community-dwelling older people [20]. In this Dutch cohort study with 2-year follow-up including 2420 community-dwelling older people, the area under the receiver operating characteristic curve (AUROC) for predicting mortality using the TFI was 0.620 [20]. Previous studies that compared alternative modeling techniques for predicting survival made use of pseudovalues [21,22]. In this study, we focused on 7-year mortality.

Methods

Study Population and Data Collection

In June 2008, the TFI was sent to a sample of 1154 community-dwelling older people aged 75 years and older randomly drawn from the register of the municipality in Roosendaal, a town of 78,000 inhabitants in the Netherlands. A total of 484 participants completed the questionnaire (41.94% response rate), which, complementary to the TFI, also contained measures for assessing quality of life and disability [23,24]. As in a previous study, the data from 5 participants were left out of the analyses as they had too many omissions, leaving a data set of 479 participants [23].

Measures

Frailty

The TFI contained 15 components of frailty distributed over physical, psychological, and social frailty. The components of physical frailty included the following: physically unhealthy, unexplained weight loss, difficulty in walking, difficulty in maintaining balance, poor hearing, poor vision, lack of strength in the hands, and physical tiredness. Psychological frailty consisted of problems with memory, feeling down, feeling nervous or anxious, and being unable to cope with problems. Social frailty included living alone, lack of social relations, and lack of social support. For the exact content and the scoring of the TFI, we refer to a previous study [13].

Mortality

In August 2015, the municipality of Roosendaal provided the mortality dates of the participants who completed the questionnaire in 2008. With these dates, 7-year mortality was defined.

Data and Data Imputation

For the modeling, we used the data set (N=479) with the 15 frailty components, gender ("male" or "female"), and the dichotomous transformed age variable ("≤80" or ">80" years) as predictor variables and 7-year mortality ("alive" or "dead") as the outcome variable. We imputed data for the missing values using the MICE (Multivariate Imputation by Chained Equations) package (m=5 and methods="logreg") in R software (version 3.4.4; The R Foundation) [25]. The first imputed data set was used for the modeling.

Modeling Techniques

Overview

We compared eight modeling techniques to predict 7-year mortality: (1) LR, (2) least absolute shrinkage and selection operator (LASSO), (3) SVM, (4) NN, (5) recursive partitioning (RP), (6) RF, (7) hill-climbing (HC) Bayesian network, and (8) naïve Bayes (NB) network.

Here, we list the main characteristics of the evaluated modeling techniques, based on the work of several authors [2,3,26-30] and an earlier publication of the first author [31].

XSL•FO
RenderX

Logistic Regression

LR is a type of regression analysis that is often used in medical research to model the probability of a dichotomous endpoint using a linear function of the predictors. Predictor variables may be either continuous or categorical. LR uses a logistic transformation to calculate the probability of a dichotomous outcome. Regression coefficients were estimated by maximum likelihood [31].

Least Absolute Shrinkage and Selection Operator

LASSO is quite similar to linear regression and LR, but it adds a penalty for nonzero regression coefficients using the sum of their absolute values. As a result, small regression coefficients are set to zero. Regression coefficients were estimated by maximum likelihood [31].

Support Vector Machine

An SVM performs classification tasks by constructing hyperplanes with a margin in a multidimensional space that separates cases from different classes. An SVM can perform a nonlinear classification or regression task using different kernels (ie, radial, linear, and polynomial). The tuning parameters for SVMs are the C parameter (cost), which regulates the margin width, and the gamma parameter for the kernel calculation. SVM claims to be a robust classification and regression technique that maximizes the predictive accuracy of a model without overfitting the training data. SVM may be particularly suited to analyze data with large numbers of predictor variables [31].

Neural Network

An NN simulates a large number of interconnected simple processing units that are arranged in layers. There are three parts in an NN: an input layer, with units representing the predictor variables; one or more hidden layers; and an output layer, with a unit representing the endpoint. The units are connected with varying connection strengths or weights. Input data are presented to the input layer, and values are propagated from there to the next layer. Then, a prediction is delivered from the output layer. The NN learns by examining individual records, generating a prediction for each record and making adjustments to the weights whenever it makes an incorrect prediction. The adjustments are based on the gradient descent algorithm to minimize the prediction error. This process is repeated many times, and the NN continues to improve its predictions until the magnitude of the gradient is less than a certain threshold (eg, 0.00005). Once trained, the NN can be applied to new records for which the endpoint is unknown. The crucial parameters of an NN are the size parameter (ie, number of units in the layer) and the decay parameter, which penalizes large weights in the model to avoid overfitting [31].

Recursive Partitioning

RP is a modeling technique that uses RP to split the training records into segments with similar endpoint values. The modeling starts by examining the input variables to find the best split, measured by the reduction in an impurity index that results from the split. The split defines two subgroups, each of which is subsequently split into two further subgroups and so on, until a stopping criterion is met. The commonly used parameter for RP is the cp parameter (cost complexity factor). A cp value of 0.001, for example, regulates that a split must decrease the overall lack of fit by a factor of 0.001 [31].

Random Forest

RF is an ensemble classifier that consists of many decision trees. In case of classification, RF outputs the class that is the mode among the classes from individual trees. In case of regression, RF outputs the value that is the mean of the values output from individual trees. Each tree is constructed using a bootstrap sample from the original data. A tree is grown by recursively partitioning the bootstrap sample based on optimization of a split rule. In regression problems, the split rule is based on minimizing the mean squared error, whereas in classification problems, the Gini index is commonly used. At each split, a subset of candidate variables are tested for the split rule optimization, similar to RP modeling. For prediction, a new sample is pushed down the tree. This procedure is iterated over all trees in the ensemble. Key parameters are the number of trees and the number of candidate variables [31].

Hill-Climbing Bayesian Network

A Bayesian network is a mathematical construct that compactly represents a joint probability distribution among a set of variables. Bayesian networks are frequently employed for modeling domain knowledge in decision support systems, particularly in medicine. Learning Bayesian networks is connected with variable selection for classification and has been used to design algorithms that optimally solve the problem under certain conditions. The HC Bayesian network is a score-based search algorithm to learn a Bayesian network structure with a sparse set of variables [32].

Naïve Bayes Network

The NB model is technically a special case of a Bayesian network. The NB model assumes that all the features are conditionally independent of each other and that, therefore, the Bayesian rule for probability can be applied. Usually this independence assumption works well for most cases, even if in actuality they are not really independent [32].

Analysis

For all analyses, we used R (version 3.4.4; The R Foundation) [33].

Statistics

We used counts and percentages to describe the baseline characteristics of the participants. The chi-square test was used to compare dichotomous variables. A *P* value of less than .05 was considered significant. Cramer V, a statistic derived from the chi-square value, was used as an association measure: values toward zero indicate weak association and values toward 1 indicate strong association. The predictive performance of the models was measured using the AUROC. An AUROC greater than 0.700 was considered as an indication of good predictive performance [3].

Relative Importance of the Predictor Variables

The relative importance of a predictor variable in a model was calculated using the Permutation Feature Importance algorithm with 1000 repetitions [34,35]. We used the decrease in median apparent AUROC as the measure for ranking the relative importance of a predictor variable.

Bootstrap Validation of the Models

Each model was validated using the bootstrap validation procedure as proposed by Efron and Tibshirani [36]. Here, we describe the bootstrap validation procedure. First, a model was developed on the original data set, and the AUROC of that model for the original data set was calculated (ie the apparent AUROC). Then, a sample with replacement was drawn from the original data set with a size equal to the size of the original data set. This sample was called the bootstrap sample. For this bootstrap sample, the model was developed again, and the AUROC for that bootstrap sample was calculated (ie, the developed AUROC). This model was then applied to the original data set and the AUROC was calculated (ie, the validated AUROC). The difference between the developed AUROC and the validated AUROC is defined as the optimism of the model. By subtracting this optimism from the apparent AUROC, we obtain the corrected AUROC. This process was repeated 100 times.

Ethics Approval and Consent to Participate

All procedures performed in studies involving human participants were in accordance with the ethical standards of the institutional and national research committee, and with the 1964 Declaration of Helsinki and its later amendments or comparable ethical standards. For this study, medical ethics approval was not necessary because particular treatments or interventions were not offered or withheld from respondents. Moreover, the integrity of the respondents was not encroached upon as a consequence of participating in this study, which is the main criterion in medical-ethical procedures in the Netherlands [37]. Informed consent related to details of the study and maintaining confidentiality was observed.

Results

Participant Characteristics and Variable Association

Table 1 presents the descriptive statistics and the univariate P values of the chi-square test for the participants at baseline in relation to 7-year mortality. Five predictor variables (ie, gender, poor hearing, poor vision, feeling down, and living alone) showed univariate P values equal to or greater than .05. Three of these predictor variables (ie, poor hearing, poor vision, and living alone) had P values equal to or greater than .20.

Table 1. Participant characteristics.

Characteristic (category)	Alive (n=317), n (%)	Dead (n=162), n (%)	P value[a]
Gender (male)	130 (41.0)	77 (47.5)	.17
Age (>80 years)	119 (37.5)	85 (52.5)	.002
Physically unhealthy (yes)	71 (22.4)	70 (43.2)	<.001
Unexplained weight loss (yes)	15 (4.7)	21 (13.0)	.001
Difficulty in walking (yes)	121 (38.2)	110 (67.9)	<.001
Difficulty in maintaining balance (yes)	86 (27.1)	84 (51.9)	<.001
Poor hearing (yes)	110 (34.7)	65 (40.1)	.24
Poor vision (yes)	65 (20.5)	38 (23.5)	.46
Lack of strength in the hands (yes)	96 (30.3)	68 (42.0)	.01
Physical tiredness (yes)	120 (37.9)	98 (60.5)	<.001
Problems with memory (yes)	21 (6.6)	25 (15.4)	.002
Feeling down (yes)	121 (38.2)	72 (44.4)	.19
Feeling nervous or anxious (yes)	87 (27.4)	61 (37.7)	.02
Unable to cope with problems (yes)	42 (13.2)	34 (21.0)	.03
Living alone (yes)	154 (48.6)	75 (46.3)	.64
Lack of social relations (yes)	174 (54.9)	108 (66.7)	.01
Lack of social support (yes)	44 (13.9)	34 (21.0)	.046

[a]Univariate P values were based on the chi-square test for the participants at baseline in relation to 7-year mortality.

A priori, we could assume that the predictor variables listed in Table 1 have no association. Figure 1 visualizes the association of the predictor variables with each other and with the outcome variable based on Cramer V, as described in the Statistics section. For example, there are strong associations between "difficulty in walking" and "difficulty in maintaining balance" and between "feeling anxious or nervous" and "feeling down."

Figure 1. Strength of the associations between the predictor variables (darker colour indicates stronger association).
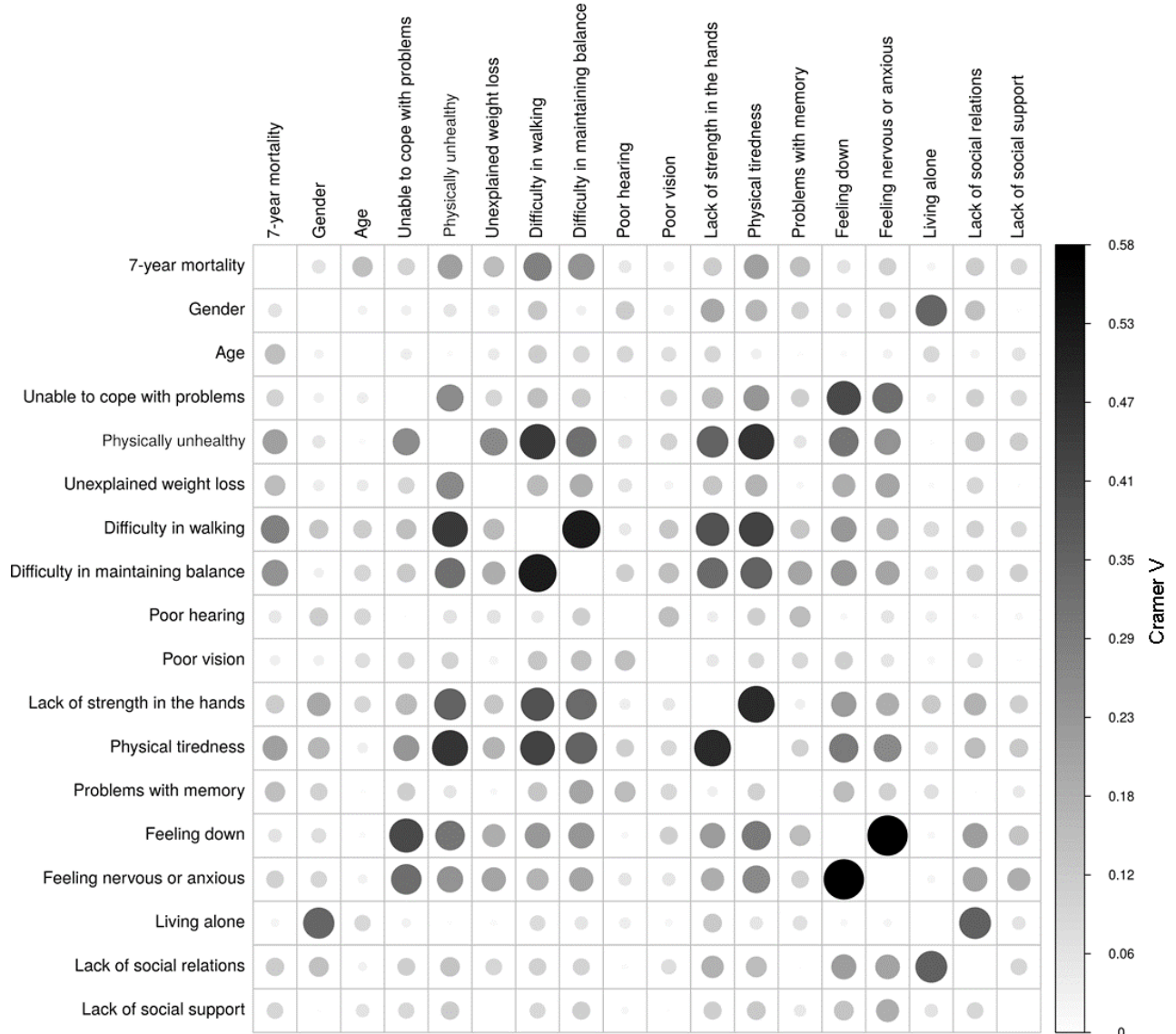
Prediction of 7-Year Mortality by the 15 Frailty Components, Gender, and Age

We applied each modeling technique, as mentioned in the Modeling Techniques section, to the data set mentioned in the Measures section and validated the models with bootstrapping (100 repetitions) as described in the Analysis section. Table 2 presents the performance characteristics of the models. The corrected AUROC values varied from 0.605 for RP to 0.812 for NN. The optimism of the NN model was high (0.156). The optimism of the RF model showed a 95% CI containing a value of zero, indicating that the RF model was not overfitted. However, the performance of the RF model was low (apparent AUROC=0.665).

Table 2. Performance characteristics of the models.

Model	Apparent AUROC[a,b]	Developed AUROC[c], mean (95% CI)	Validated AUROC[d], mean (95% CI)	Optimism[e], mean (95% CI)	Corrected AUROC[f]
Logistic regression	0.743	0.765 (0.723 to 0.804)	0.721 (0.694 to 0.735)	0.045 (0.006 to 0.084)	0.698
LASSO[g]	0.742	0.762 (0.717 to 0.799)	0.720 (0.700 to 0.733)	0.043 (0.006 to 0.084)	0.699
SVM[h]	0.764	0.804 (0.771 to 0.837)	0.745 (0.729 to 0.763)	0.059 (0.020 to 0.089)	0.705
Neural network	0.967	0.989 (0.974 to 0.998)	0.834 (0.793 to 0.868)	0.156 (0.123 to 0.197)	0.812
Recursive partitioning	0.680	0.771 (0.711 to 0.826)	0.696 (0.643 to 0.731)	0.075 (0.034 to 0.116)	0.605
Random forest	0.665	0.867 (0.835 to 0.899)	0.873 (0.851 to 0.898)	–0.007 (–0.056 to 0.042)	0.671
HC[i] Bayesian network	0.649	0.674 (0.522 to 0.738)	0.654 (0.521 to 0.689)	0.020 (–0.009 to 0.061)	0.629
Naïve Bayes	0.704	0.717 (0.683 to 0.759)	0.704 (0.704 to 0.704)	0.014 (–0.021 to 0.055)	0.690

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

[b]The apparent AUROC is the AUROC of the model for the original data set.

[c]The developed AUROC is the AUROC of the redeveloped model on the bootstrap sample.

[d]The validated AUROC is the AUROC of the validated model.

[e]The model optimism is the difference between the developed AUROC and the validated AUROC.

[f]The corrected AUROC is the AUROC obtained by subtracting the optimism from the apparent AUROC.

[g]LASSO: least absolute shrinkage and selection operator.

[h]SVM: support vector machine.

[i]HC: hill-climbing.

Relative Importance of the Predictor Variables for the NN Model and the SVM Model

The NN model and the SVM model had corrected AUROCs above 0.700, indicating a good performance. Figure 2 shows the relative importance of the predictor variables for these models, calculated as described in the Analysis section. The depicted points correspond to the median decrease in apparent AUROC, and the boundaries of the bands illustrate the 95% CI for the decrease in apparent AUROC. The dashed line corresponds to a value of zero. If the 95% CI contains the value of zero, the predictor variable has no significant importance for the model. The predictor variables "difficulty in walking," "gender," and "difficulty in maintaining balance" had the highest relative importance in the NN model; the predictor variables "age," "feeling down," and "difficulty in walking" had the highest relative importance in the SVM model. For the relative importance of the predictor variables in the other models, we refer to Figures S1-S3 in Multimedia Appendix 1.

Figure 2. Median decrease in apparent AUROC and 95% CI (whiskers) for the neural network model (A) and the support vector machine model (B). AUROC: area under the receiver operating characteristic curve.
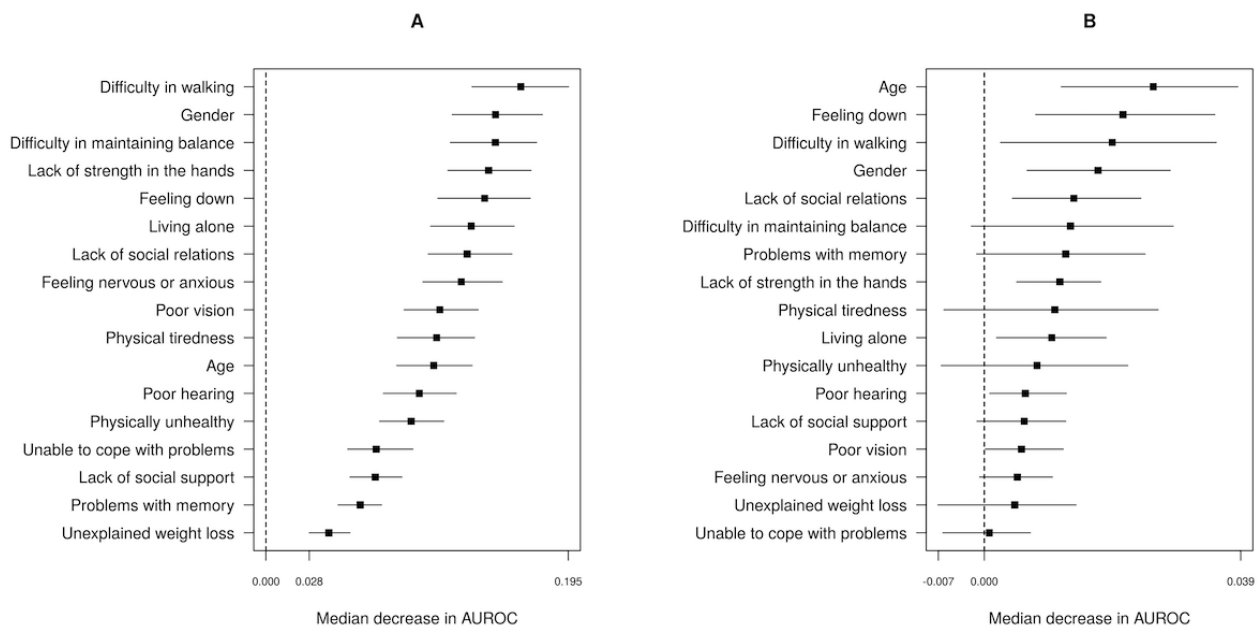
Discussion

Principal Findings

Many studies have observed that frailty is associated with mortality among community-dwelling older people [38]. To date, only one study used the original version of the TFI for the prediction of mortality among Dutch community-dwelling older people, using a 2-year follow-up [20].

The aim of this study was to determine the best modeling technique for predicting mortality in a Dutch sample of 479 community-dwelling older people with a 7-year follow-up by assessing frailty with the TFI. We compared eight modeling techniques to develop prediction models. The classical approach for developing a prediction model for a dichotomous outcome is to use the LR technique or the penalized version, LASSO. Both techniques are based on a linear combination of the predictor variables (see Modeling Techniques section). The other evaluated techniques are able to capture nonlinearity and can deal with interaction of the predictor variables [39].

Of the 15 components of the TFI, three had *P* values equal to or greater than .20 (ie, poor hearing, poor vision, and living alone); normally, these variables would not be included in a multivariate analysis. However, removing these components from the TFI on the basis of this study is not recommended. The inclusion of sensory difficulties in a screening instrument such as the TFI has major consequences in terms of the prevalence and prediction of adverse outcomes (eg, hospitalization) [40]. Therefore, for all techniques, we used all 15 components of the TFI; we also added "gender" and "age" as predictor variables.

The simplest way to construct a prediction model is to calculate the sum score of the TFI components, adding 1 if the participant is "male" and adding 1 again if the participant is ">80 years." Therefore, the maximum sum score is 17. The apparent AUROC

for this sum score model in predicting mortality was 0.680. The algorithm of the LR modeling technique led to a model with an apparent AUROC of 0.743 in predicting mortality. The LASSO model had an apparent AUROC of 0.742, with only the following predictors: "age," "physically unhealthy," "difficulty in walking," "difficulty in maintaining balance," and "physical tiredness." These results show that applying algorithms paid off above using just the simple approach.

LR and LASSO are regression-based techniques. An SVM is a modern, advanced modeling technique that is able to discriminate between the categories "alive" and "dead" using high-dimensional hyperplanes to separate them. The corrected AUROC of the SVM model was 0.705 and the optimism was 0.059.

The NN model showed the highest apparent and corrected AUROCs. However, the optimism of the NN model was 0.156. This and the fact that an NN model has a black box character makes the application of an NN model unattractive in predicting mortality in our study.

We calculated the relative importance of the predictors in the NN model as well as in the SVM model. It is obvious that the top three important variables differed for both models. However, the predictor variable "difficulty in walking" was present in the top three of both models. This was also the case with the other six models. In general, each model has its own ranking of important variables due to the underlying algorithm [21].

Models provided by the RP modeling technique are considered attractive in a medical setting because they show a decision tree. In our study, the RP model performed poorly (corrected AUROC=0.605). The RF modeling technique is attractive because it claims to provide models without overfitting [26]. This is in line with our study because the 95% CI for bootstrap validation for the optimism was –0.056 to 0.042, indicating that the optimism does not differ significantly from zero. The

performance of the RF model was also somewhat poor (corrected AUROC=0.671). However, the RF modeling technique is considered as an obvious improvement over the RP modeling technique [41,42]. It is, hence, remarkable that the RP modeling technique has, until recently, been advocated for as the preferred modeling technique for prediction in some disease areas, such as trauma [4].

Bayesian networks, with their underlying algorithms, are especially suited for capturing and reasoning with uncertainty. They have been applied in biomedicine and health care for more than a decade now and are still gaining in popularity. Bayesian networks are used in clinical epidemiology for the construction of disease prediction models and within bioinformatics for the interpretation of microarray gene expression data, for instance [43]. In our study, we evaluated two Bayesian network algorithms, HC Bayesian network and NB, for the prediction of 7-year mortality. The HC Bayesian network and NB algorithms showed corrected AUROCs of 0.629 and 0.690, respectively. The NB algorithm used all predictor variables, whereas the HC Bayesian network algorithm was developed to determine a sparse set of predictor variables. For our data set, the HC Bayesian network algorithm only used the predictor variable "difficulty in walking" for the prediction of 7-year mortality.

The internal validation of the models was done using bootstrapping with 100 repetitions to get insight into the amount of optimism. Other examples of internal validation techniques are split-sample and cross-validation techniques [44]. While the interest in the development, validation, and clinical application of prediction models is increasing, a recent systematic review showed that only a quarter of the studies reported prediction models with internal as well as external validation [45,46]. External validation aims to address the performance of a prediction model in a different but plausibly related data set, which still represents the underlying domain. This validation step is widely considered necessary before implementing a developed prediction model in practice [47,48]. We support this notion, and we strongly suggest validating the developed models in our study in the data sets that were used in other studies [16-20].

A number of limitations of this study should be addressed. First, our sample consisted exclusively of people living independently in the municipality of Roosendaal. Therefore, the generalizability of the findings can be questioned. Second, the TFI is a frailty instrument using self-reported data, so frailty is subjectively assessed. However, the construct validity of the TFI has been determined in detail using objective measurements [13]. Third, we used default settings for the modeling techniques. This holds for LR and LASSO as well as for the modern methods where various specific parameters might be fine-tuned to the development setting [1,3,42]. Further tuning of parameters to specific issues in a particular development data set might obviously improve the apparent performance, but we doubt that substantial improvement would be achieved in the validated external performance.

Conclusions

In conclusion, this study has shown that the NN and SVM models outperformed the other six models (corrected AUROCs>0.700). Because of the high optimism of the NN model, we prefer the SVM model for predicting mortality among community-dwelling older people using the 15 components of the TFI, with the addition of "gender" and "age." Furthermore, external validation is a necessary step before applying the prediction models in a new setting.

Data Availability

The data set used and analyzed during this study is available from the corresponding author on reasonable request.

Authors' Contributions

TvdP and RG wrote the main manuscript text. TvdP prepared all figures and all tables and performed all analyses. TvdP and RG reviewed the manuscript.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Relative importance of the predictor variables for the models.
[DOCX File , 415 KB - medinform_v10i3e31480_app1.docx]

References

1. Hastie T, Tibshirani R, Friedman J. The Elements of Statistical Learning: Data Mining, Inference, and Prediction. 2nd edition. New York, NY: Springer Science+Business Media; 2017.

2.   Harrell Jr FE. Regression Modeling Strategies: With Applications to Linear Models, Logistic and Ordinal Regression, and Survival Analysis. 2nd edition. Cham, Switzerland: Springer International Publishing; 2015.

3.   Steyerberg EW. Clinical Prediction Models: A Practical Approach to Development, Validation, and Updating. 2nd edition. Cham, Switzerland: Springer Nature Switzerland AG; 2019.

4.   Young NH, Andrews PJD. Developing a prognostic model for traumatic brain injury--A missed opportunity? PLoS Med 2008 Aug 05;5(8):e168 [FREE Full text] [doi: 10.1371/journal.pmed.0050168] [Medline: 18684010]

5.   Biganzoli E, Boracchi P, Mariani L, Marubini E. Feed forward neural networks for the analysis of censored survival data: A partial logistic regression approach. Stat Med 1998 May 30;17(10):1169-1186. [doi: 10.1002/(sici)1097-0258(19980530)17:10<1169::aid-sim796>3.0.co;2-d]

6.   Ishwaran H, Kogalur UB, Blackstone EH, Lauer MS. Random survival forests. Ann Appl Stat 2008 Sep 1;2(3):841-860. [doi: 10.1214/08-aoas169]

7.   Lee YJ, Mangasarian OL, Wolberg WH. Breast cancer survival and chemotherapy: A support vector machine analysis. In: Proceedings of the DIMACS Workshop: Discrete Mathematical Problems With Medical Applications. 2000 Presented at: DIMACS Workshop: Discrete Mathematical Problems With Medical Applications; December 8-10, 1999; Piscataway, NJ p. 1-10 URL: https://dsmilab.github.io/Yuh-Jye-Lee/assets/file/publications/book_chapter/B3_Breast%20Cancer%20Survival%20and%20Chemotherapy%20A%20Support%20Vector%20Machine%20Analysis.pdf [doi: 10.1090/dimacs/055/01]

8.   Ture M, Tokatli F, Kurt Omurlu I. The comparisons of prognostic indexes using data mining techniques and Cox regression analysis in the breast cancer data. Expert Syst Appl 2009 May;36(4):8247-8254. [doi: 10.1016/j.eswa.2008.10.014]

9.   Fried LP, Tangen CM, Walston J, Newman AB, Hirsch C, Gottdiener J, Cardiovascular Health Study Collaborative Research Group. Frailty in older adults: Evidence for a phenotype. J Gerontol A Biol Sci Med Sci 2001 Mar;56(3):M146-M156. [doi: 10.1093/gerona/56.3.m146] [Medline: 11253156]

10. Kojima G, Iliffe S, Jivraj S, Walters K. Association between frailty and quality of life among community-dwelling older people: A systematic review and meta-analysis. J Epidemiol Community Health 2016 Jul;70(7):716-721. [doi: 10.1136/jech-2015-206717] [Medline: 26783304]

11. Vermeiren S, Vella-Azzopardi R, Beckwée D, Habbig A, Scafoglieri A, Jansen B, Gerontopole Brussels Study group. Frailty and the prediction of negative health outcomes: A meta-analysis. J Am Med Dir Assoc 2016 Dec 01;17(12):1163.e1-1163.e17. [doi: 10.1016/j.jamda.2016.09.010] [Medline: 27886869]

12. Gobbens RJJ, Luijkx KG, Wijnen-Sponselee MT, Schols JMGA. Towards an integral conceptual model of frailty. J Nutr Health Aging 2010 Mar;14(3):175-181. [doi: 10.1007/s12603-010-0045-6] [Medline: 20191249]

13. Gobbens RJ, van Assen MA, Luijkx KG, Wijnen-Sponselee MT, Schols JM. The Tilburg Frailty Indicator: Psychometric properties. J Am Med Dir Assoc 2010 Jun;11(5):344-355. [doi: 10.1016/j.jamda.2009.11.003] [Medline: 20511102]

14. Gobbens RJ, Luijkx KG, Wijnen-Sponselee MT, Schols JM. In search of an integral conceptual definition of frailty: Opinions of experts. J Am Med Dir Assoc 2010 Jun;11(5):338-343. [doi: 10.1016/j.jamda.2009.09.015] [Medline: 20511101]

15. Gobbens RJ, Uchmanowicz I. Assessing frailty with the Tilburg Frailty Indicator (TFI): A review of reliability and validity. Clin Interv Aging 2021 May;16:863-875. [doi: 10.2147/cia.s298191]

16. Santiago LM, Gobbens RJ, van Assen MA, Carmo CN, Ferreira DB, Mattos IE. Predictive validity of the Brazilian version of the Tilburg Frailty Indicator for adverse health outcomes in older adults. Arch Gerontol Geriatr 2018;76:114-119. [doi: 10.1016/j.archger.2018.02.013] [Medline: 29494871]

17. Andreasen J, Aadahl M, Sørensen EE, Eriksen HH, Lund H, Overvad K. Associations and predictions of readmission or death in acutely admitted older medical patients using self-reported frailty and functional measures. A Danish cohort study. Arch Gerontol Geriatr 2018;76:65-72. [doi: 10.1016/j.archger.2018.01.013] [Medline: 29462759]

18. Huisman M, Deeg D. The course of frailty. In: van Campen C, editor. Frail Older Persons in the Netherlands. The Hague, the Netherlands: The Netherlands Institute for Social Research; Sep 2011:83-90.

19. Theou O, Brothers TD, Mitnitski A, Rockwood K. Operationalization of frailty using eight commonly used scales and comparison of their ability to predict all-cause mortality. J Am Geriatr Soc 2013 Sep;61(9):1537-1551. [doi: 10.1111/jgs.12420] [Medline: 24028357]

20. Op Het Veld LPM, Beurskens AJHM, de Vet HCW, van Kuijk SMJ, Hajema K, Kempen GIJM, et al. The ability of four frailty screening instruments to predict mortality, hospitalization and dependency in (instrumental) activities of daily living. Eur J Ageing 2019 Sep;16(3):387-394 [FREE Full text] [doi: 10.1007/s10433-019-00502-4] [Medline: 31543731]

21. van der Ploeg T, Datema F, Baatenburg de Jong R, Steyerberg EW. Prediction of survival with alternative modeling techniques using pseudo values. PLoS One 2014;9(6):e100234 [FREE Full text] [doi: 10.1371/journal.pone.0100234] [Medline: 24950066]

22. Klein JP, Gerster M, Andersen PK, Tarima S, Perme MP. SAS and R functions to compute pseudo-values for censored data regression. Comput Methods Programs Biomed 2008 Mar;89(3):289-300 [FREE Full text] [doi: 10.1016/j.cmpb.2007.11.017] [Medline: 18199521]

23. Gobbens RJJ, van Assen MALM, Luijkx KG, Schols JMGA. The predictive validity of the Tilburg Frailty Indicator: Disability, health care utilization, and quality of life in a population at risk. Gerontologist 2012 Oct;52(5):619-631. [doi: 10.1093/geront/gnr135] [Medline: 22217462]

XSL•FO

RenderX

24.   Gobbens RJ, van Assen MA, Luijkx KG, Wijnen-Sponselee MT, Schols JM. Determinants of frailty. J Am Med Dir Assoc
       2010 Jun;11(5):356-364. [doi: 10.1016/j.jamda.2009.11.008] [Medline: 20511103]

25.   van Buuren S, Groothuis-Oudshoorn K. mice: Multivariate imputation by chained equations in R. J Stat Softw
       2011;45(3):1-67. [doi: 10.18637/jss.v045.i03]

26.   Breiman L, Friedman JH, Olshen RA, Stone CJ. Classification and Regression Trees. Boca Raton, FL: Chapman & Hall/CRC;
       1984:1-368.

27.   Tufféry S. Data Mining and Statistics for Decision Making. Hoboken, NJ: John Wiley & Sons; 2011:1-720.

28.   Cortes C, Vapnik V. Support-vector networks. Mach Learn 1995 Sep;20(3):273-297. [doi: 10.1007/bf00994018]

29.   Tsamardinos I, Aliferis CF. Towards principled feature selection: Relevancy, filters and wrappers. In: Proceedings of the
       9th International Workshop on Artificial Intelligence and Statistics. 2003 Presented at: The 9th International Workshop on
       Artificial Intelligence and Statistics; January 3-6, 2003; Key West, FL p. 300-307 URL: http://proceedings.mlr.press/r4/
       tsamardinos03a/tsamardinos03a.pdf

30.   Aliferis CF, Tsamardinos I, Statnikov AR, Brown LE. Causal Explorer: A causal probabilistic network learning toolkit for
       biomedical discovery. In: Proceedings of the International Conference on Mathematics and Engineering Techniques in
       Medicine and Biological Sciences. 2003 Presented at: International Conference on Mathematics and Engineering Techniques
       in Medicine and Biological Sciences; June 23-26, 2003; Las Vegas, NV p. 371-376.

31.   van der Ploeg T, Steyerberg EW. Feature selection and validated predictive performance in the domain of Legionella
       pneumophila: A comparative study. BMC Res Notes 2016 Mar 08;9:147 [FREE Full text] [doi: 10.1186/s13104-016-1945-2]
       [Medline: 26951763]

32.   Lowd D, Domingos P. Naive Bayes models for probability estimation. In: Proceedings of the 22nd International Conference
       on Machine Learning. 2005 Presented at: The 22nd International Conference on Machine Learning; August 7-11, 2005;
       Bonn, Germany p. 529-536. [doi: 10.1145/1102351.1102418]

33.   R Core Team. The R Project for Statistical Computing. Vienna, Austria: The R Foundation URL: https://www.r-project.org/
       [accessed 2022-03-14]

34.   Breiman L. Random forests. Mach Learn 2001 Oct;45:5-32 [FREE Full text] [doi: 10.1023/A:1010933404324]

35.   Fisher A, Rudin C, Dominici F. All models are wrong, but many are useful: Learning a variable's importance by studying
       an entire class of prediction models simultaneously. J Mach Learn Res 2019;20:177 [FREE Full text] [Medline: 34335110]

36.   Efron B, Tibshirani RJ. An Introduction to the Bootstrap. Boca Raton, FL: Chapman & Hall/CRC; 1994:1-456.

37.   Your research: Is it subject to the WMO or not? Central Committee on Research Involving Human Subjects. URL: https:/
       /english.ccmo.nl/investigators/legal-framework-for-medical-scientific-research/your-research-is-it-subject-to-the-wmo-or-not
       [accessed 2022-03-14]

38.   Shamliyan T, Talley KM, Ramakrishnan R, Kane RL. Association of frailty with survival: A systematic literature review.
       Ageing Res Rev 2013 Mar;12(2):719-736. [doi: 10.1016/j.arr.2012.03.001] [Medline: 22426304]

39.   van der Ploeg T. Prediction of Medical Outcomes with Modern Modelling Techniques [doctoral thesis]. Rotterdam, the
       Netherlands: Erasmus University Rotterdam; 2017. URL: https://repub.eur.nl/pub/95059 [accessed 2022-03-14]

40.   Linard M, Herr M, Aegerter P, Czernichow S, Goldberg M, Zins M, et al. Should sensory impairment be considered in
       frailty assessment? A study in the GAZEL cohort. J Nutr Health Aging 2016;20(7):714-721. [doi:
       10.1007/s12603-015-0651-4] [Medline: 27499304]

41.   van der Ploeg T, Nieboer D, Steyerberg EW. Modern modeling techniques had limited external validity in predicting
       mortality from traumatic brain injury. J Clin Epidemiol 2016 Oct;78:83-89. [doi: 10.1016/j.jclinepi.2016.03.002] [Medline:
       26987507]

42.   Steyerberg EW, Eijkemans MJ, Harrell FE, Habbema JDF. Prognostic modelling with logistic regression analysis: A
       comparison of selection and estimation methods in small data sets. Stat Med 2000 Apr 30;19(8):1059-1079. [doi:
       10.1002/(sici)1097-0258(20000430)19:8<1059::aid-sim412>3.0.co;2-0]

43.   Lucas PJ, van der Gaag LC, Abu-Hanna A. Bayesian networks in biomedicine and health-care. Artif Intell Med 2004
       Mar;30(3):201-214. [doi: 10.1016/j.artmed.2003.11.001] [Medline: 15081072]

44.   Steyerberg EW, Harrell FE, Borsboom GJ, Eijkemans M, Vergouwe Y, Habbema JF. Internal validation of predictive
       models. J Clin Epidemiol 2001 Aug;54(8):774-781. [doi: 10.1016/s0895-4356(01)00341-9]

45.   Collins GS, de Groot JA, Dutton S, Omar O, Shanyinde M, Tajar A, et al. External validation of multivariable prediction
       models: A systematic review of methodological conduct and reporting. BMC Med Res Methodol 2014 Mar 19;14:40 [FREE
       Full text] [doi: 10.1186/1471-2288-14-40] [Medline: 24645774]

46.   Siontis GC, Tzoulaki I, Castaldi PJ, Ioannidis JP. External validation of new risk prediction models is infrequent and reveals
       worse prognostic discrimination. J Clin Epidemiol 2015 Jan;68(1):25-34. [doi: 10.1016/j.jclinepi.2014.09.007] [Medline:
       25441703]

47.   Bleeker S, Moll H, Steyerberg E, Donders A, Derksen-Lubsen G, Grobbee D, et al. External validation is necessary in
       prediction research: A clinical example. J Clin Epidemiol 2003 Sep;56(9):826-832. [doi: 10.1016/s0895-4356(03)00207-5]

48.   Justice AC, Covinsky KE, Berlin JA. Assessing the generalizability of prognostic information. Ann Intern Med 1999 Mar
       16;130(6):515-524. [doi: 10.7326/0003-4819-130-6-199903160-00016] [Medline: 10075620]

Abbreviations

AUROC: area under the receiver operating characteristic curve
HC: hill-climbing
LASSO: least absolute shrinkage and selection operator
LR: logistic regression
MICE: Multivariate Imputation by Chained Equations
NB: naïve Bayes
NN: neural network
RF: random forest
RP: recursive partitioning
SVM: support vector machine
TFI: Tilburg Frailty Indicator

XSL•FO
RenderX

<u>Original Paper</u>

Mining Electronic Health Records for Drugs Associated With 28-day Mortality in COVID-19: Pharmacopoeia-wide Association Study (PharmWAS)

Ivan Lerner[1,2,3], MD; Arnaud Serret-Larmande[1,2], MD; Bastien Rance[1,3], PhD; Nicolas Garcelon[3,4], PhD; Anita Burgun[1,2,3], MD, PhD; Laurent Chouchana[5], PhD, PharmD; Antoine Neuraz[1,2,3], MD, PhD

[1]Inserm, Centre de Recherche des Cordeliers, Sorbonne Université, Paris, France

[2]Informatique biomédicale, Hôpital Necker-Enfants Malades, Assistance Publique - Hôpitaux de Paris, Paris, France

[3]HeKA Team, Inria, Paris, France

[4]Inserm UMR 1163, Data Science Platform, Université de Paris, Imagine Institute, Paris, France

[5]Centre Régional de Pharmacovigilance, Service de Pharmacologie, Hôpital Cochin, Assistance Publique - Hôpitaux de Paris, Centre - Université de Paris, Paris, France

Corresponding Author:
Antoine Neuraz, MD, PhD
Inserm
Centre de Recherche des Cordeliers
Sorbonne Université
Université de Paris
15 Rue de l'École de Médecine
Paris, 75006
France
Phone: 33 01 44 27 64 82
Email: antoine.neuraz@aphp.fr

Related Article:

This is a corrected version. See correction statement: https://medinform.jmir.org/2022/4/e38505

Abstract

Background: Patients hospitalized for a given condition may be receiving other treatments for other contemporary conditions or comorbidities. The use of such observational clinical data for pharmacological hypothesis generation is appealing in the context of an emerging disease but particularly challenging due to the presence of drug indication bias.

Objective: With this study, our main objective was the development and validation of a fully data-driven pipeline that would address this challenge. Our secondary objective was to generate pharmacological hypotheses in patients with COVID-19 and demonstrate the clinical relevance of the pipeline.

Methods: We developed a pharmacopeia-wide association study (PharmWAS) pipeline inspired from the PheWAS methodology, which systematically screens for associations between the whole pharmacopeia and a clinical phenotype. First, a fully data-driven procedure based on adaptive least absolute shrinkage and selection operator (LASSO) determined drug-specific adjustment sets. Second, we computed several measures of association, including robust methods based on propensity scores (PSs) to control indication bias. Finally, we applied the Benjamini and Hochberg procedure of the false discovery rate (FDR). We applied this method in a multicenter retrospective cohort study using electronic medical records from 16 university hospitals of the Greater Paris area. We included all adult patients between 18 and 95 years old hospitalized in conventional wards for COVID-19 between February 1, 2020, and June 15, 2021. We investigated the association between drug prescription within 48 hours from admission and 28-day mortality. We validated our data-driven pipeline against a knowledge-based pipeline on 3 treatments of reference, for which experts agreed on the expected association with mortality. We then demonstrated its clinical relevance by screening all drugs prescribed in more than 100 patients to generate pharmacological hypotheses.

Results: A total of 5783 patients were included in the analysis. The median age at admission was 69.2 (IQR 56.7-81.1) years, and 3390 (58.62%) of the patients were male. The performance of our automated pipeline was comparable or better for controlling

bias than the knowledge-based adjustment set for 3 reference drugs: dexamethasone, phloroglucinol, and paracetamol. After correction for multiple testing, 4 drugs were associated with increased in-hospital mortality. Among these, diazepam and tramadol were the only ones not discarded by automated diagnostics, with adjusted odds ratios of 2.51 (95% CI 1.52-4.16, Q=.01) and 1.94 (95% CI 1.32-2.85, Q=.02), respectively.

Conclusions: Our innovative approach proved useful in generating pharmacological hypotheses in an outbreak setting, without requiring a priori knowledge of the disease. Our systematic analysis of early prescribed treatments from patients hospitalized for COVID-19 showed that diazepam and tramadol are associated with increased 28-day mortality. Whether these drugs could worsen COVID-19 needs to be further assessed.

Introduction

COVID-19 has been a global threat for public health since its emergence in China in December 2019. On July 1, 2021, more than 182 million cases of COVID-19 were reported worldwide, including more than 3.9 million deaths [1].

Multiple scientific questions have emerged over the course of the pandemic. Tremendous efforts toward finding adequate treatment options have been taken to the point that as of August 18, 2021, 2658 clinical trials were listed by the French Cochrane Centre [2]. To date, the most notable finding was that in the inflammatory phase of the disease, dexamethasone, a systemic glucocorticoid, showed a reduction in 28-day mortality among critical patients receiving respiratory support [3]. In addition, questions regularly arise regarding the safety profiles of known drugs (eg, nonsteroidal anti-inflammatory drugs [NSAIDs], angiotensin-converting enzyme [ACE] inhibitors) [4-6] or potential drug repurposing (eg, ivermectin, fluvoxamine) [7,8]. These clinical trials are motivated by in vitro efficacy of molecules [8,9], by epidemiological observations, or by both [7,10]. Furthermore, the understanding of COVID-19's physiopathology has rapidly evolved. Hence, having understood the inflammatory component of severe cases and proven the benefit of dexamethasone in patients with severe COVID-19, dozens of immunosuppressant molecules are being tested in clinical trials [2]. At the same time, high rates of venous thromboembolism in hospitalized patients have been reported, 14.1% (95% CI 11.6-16.9) compared to 2.8%-5.6% before the pandemic [11-13], which led to multiple investigations on anticoagulant treatments.

However, 2 questions can be raised in the context of an emergent disease: (1) Are there pharmacological hypotheses that were not explored due to an incomplete physiological understanding of the disease, and (2) how can we better prioritize hypotheses to improve clinical research efficiency?

This context motivated the development of a systematic and data-driven approach that could guide clinical and epidemiological research by mining routinely collected data from electronic health records (EHRs) without the necessity of a priori knowledge. For that purpose, we took inspiration from the phenome-wide association study (PheWAS) model [14-17] to derive its drug counterpart, the pharmacopeia-wide association study (PharmWAS). This methodology analyzes in a hypothesis-free approach the association of the whole set of drug exposure with the phenotypes of a given population, similarly to a PheWAS, which analyzes the association of the whole set of phenotypes with genetic variants. The idea of PharmWAS has gained popularity in recent years under different names and has been implemented under different designs [17-20]. The PharmWAS methodology was first described by Ryan et al [17] in 2013 using a self-controlled case approach to detect adverse events. A methodology based on Cox survival models was applied by Patel et al [18] to discover drugs associated with cancer risk.

The principal challenge of a PharmWAS is to control the treatment-specific indication bias for multiple treatments. For that purpose, we developed a 2-step pipeline motivated by the literature on *causal variable selection* [21-23] that we empirically validated using reference drugs. This pipeline had to be fully data driven in order to scale to a large number of drugs. Our implementation combined a multivariate regression adjustment model and 2 PS-based methods: PS weighting and matching [24-27]. Each method represented different trade-offs between precision of the estimation and robustness to confounding. The rationale was not to report exact treatment effects, which would require domain expert knowledge supporting strong assumptions for a large set of drugs, and necessitate the strict respect of causal inference assumptions: no unmeasured confounders (exchangeability), every patient having a nonzero probability of being treated or not (positivity), and well specified models [26]. Instead, our goal was to generate pharmacological hypotheses, and we assumed that the combination of these models would reduce false-positive findings caused by indication bias.

Our main objective was to develop and validate a fully data-driven pipeline addressing these challenges. Our secondary objective was to generate pharmacological hypotheses, whether to highlight potential candidates for COVID-19 treatment or prevention or to highlight drugs worsening the condition of patients with COVID-19. To that end, we screened for associations between all drugs prescribed in the first 48 hours after admission and 28-day mortality in adults hospitalized for COVID-19 in a conventional ward.

XSL•FO

RenderX

Methods

Study Design and Data Sources

We performed a multicentric retrospective study using the Entrepôt de Données de Santé (EDS)-COVID database, developed upon the Assistance Publique - Hôpitaux de Paris (AP-HP) clinical data warehouse (CDW), regrouping data from 39 different sites in the Greater Paris area [28]. We used 4 types of data in this study: (1) medicoadministrative data, including diagnosis codes recorded using the *International Classification of Diseases, 10th edition* (ICD-10); (2) laboratory results from admission; (3) physiological measurements (eg, blood pressure) at admission; and (4) all medical text reports associated with inpatient stays.

Population

Selection of the study population was performed according to the following criteria: (1) first admission with an ICD-10 code of U07.1 (COVID-19), (2) age at admission between 18 and 95 years, (3) hospitalization in a conventional ward for at least 48 hours, (4) hospitalization in an AP-HP site uploading drug information to the CDW, and (5) exclusion of patients who initiated palliative therapy within 48 hours (Figure 1). The study time frame spanned from February 1, 2020, to June 15, 2021.

Figure 1. Flowchart and the PharmWAS pipeline. CDW: clinical data warehouse; ICD-10: International Classification of Diseases, 10th edition; LASSO: least absolute shrinkage and selection operator; PharmWAS: pharmacopeia-wide association study; PS: propensity score.
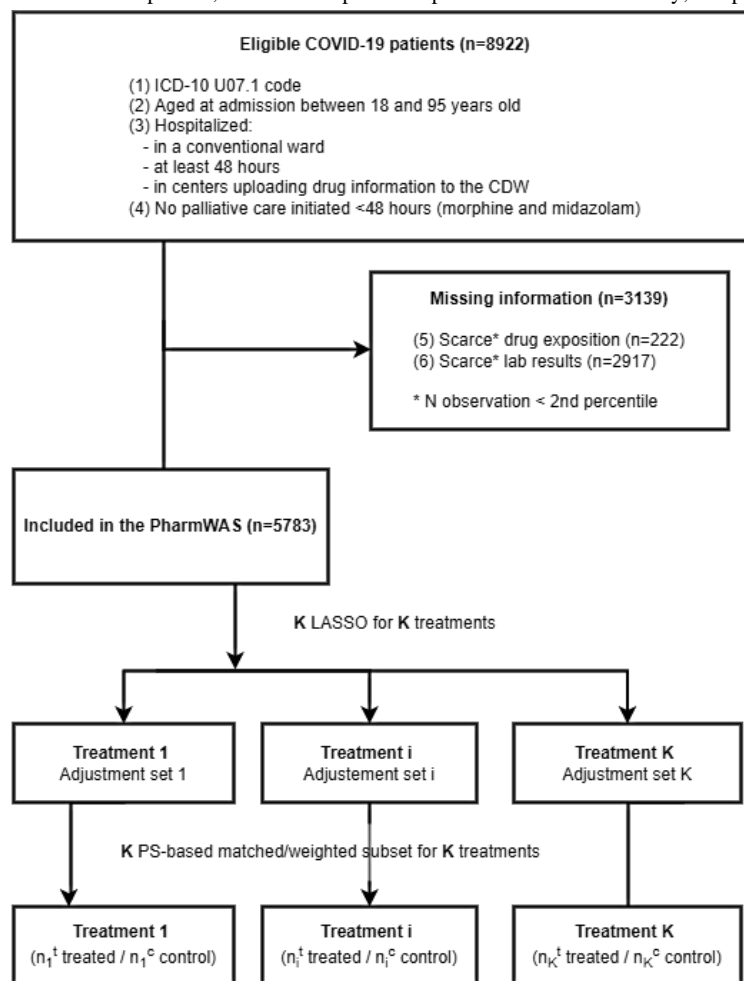
Drug Exposure and Clinical Endpoint Definition

We extracted each patient's drug exposure status from the CDW corresponding to the first 48 hours after the patient was admitted for COVID-19. A code from the anatomical therapeutic chemical (ATC) classification [29] was assigned to patients with at least 1 corresponding drug regardless of the dose. We restricted the analysis to ATC level 5 codes that were present for a minimum of 100 patients. In the following sections, we use the term *drugs* to refer to these codes. The outcome was defined as all-cause 28-day mortality, with patients discharged alive before 28 days assumed to be alive at 28 days.

Adjustment Covariate Definition

First, we included the ICD-10 codes from the current inpatient stay, restricted to chronic diseases. These codes were then grouped into broader categories using the first 3 characters and the first 2 characters of the ICD-10 system [16]. Second, we considered all laboratory results and physiological measurements. For covariates measured in at least 10% of patients, we kept only the first observation within 48 hours after admission. For covariates measured in at least 5% of patients, we kept indicator variables of the measure (1 if measured, 0 if not measured). The BMI was extracted from clinical reports using regular expressions. Finally, we added some feature

engineered variables, accounting for the study period by using quintiles of the time since study initiation and quintiles of the number of measurements by source (eg, number of lab results). All continuous variables were winsorized at 2nd and 98th percentiles to account for outliers.

Pharmacopeia-wide Association Study Pipeline

The core principle is to test for the association between each drug exposure and the outcome, controlling for covariates, given an adjustment set. This analysis is repeated n times per outcome, with n being the number of drug exposures. The results of the n tests (P values) obtained from this process are subsequently corrected to consider the multiple testing.

In the first step of the pipeline, we determined adjustment sets for every drug exposure, given the set of all possible pretreatment covariates. Using an adaptive LASSO procedure [30], we kept covariates associated with each drug from the subset of covariates associated with the outcome, after cross-validation of the model's deviance. We included for each continuous covariate 3 possible forms: square root transformation, log transformation, and discretization in quintiles.

In the second step, we computed the conditional odds ratio (OR) between the drug and the outcome in a multivariate logistic regression model, given the selected covariates. In addition, we produced 2 supplementary measures of association based on PSs as secondary analyses, namely the marginal OR on the matched population and the marginal OR on the population after inverse probability weighting (IPW), restricted to the "empirical equipoise region" (EER, ie, after trimming) [27]. The EER is defined to approximate the region of *clinical equipoise*, among which uncertainty among treatment options is strong enough, so that prescribers' preference drives the prescription instead of only patients' characteristics [27]. PS models were fitted using multivariate logistic regression. The matching procedure was implemented with a case-control ratio of 1:4 and a caliper of 0.2 SD of the logit of the PS [31]. With IPW, the cohort was resampled by weighting each individual "i" with a weight that was based on its estimated stabilized PS π_i (preference score) [27]. Stabilized PS or preference scores were PS-corrected for prevalence (logit of the preference score = logit of the PS –logit of drug prevalence). Treated individuals were then weighted by $1/\pi_i$, and controls were weighted by $1/(1 - \pi_i)$. Patients with stabilized PS outside the EER (ie, the stabilized PS interval of 0.3-0.7) were discarded [27]. Finally, both PS-based methods allowed the generation of automated diagnostics to assess the validity of the estimates: first, the residual imbalance in covariates, which we reported as the fraction of balanced covariates (FBC; ie, covariates with absolute standardized mean difference [ASMD] between treatment groups<0.1) [32], and second, the fraction of exposed population (FEP) remaining after applying the caliper in the matched subset or within the EER in the trimmed subset. Alpha risk inflation caused by multiple testing was addressed following the Benjamini and Hochberg procedure of the FDR (Q=.05), and P values for the OR were corrected accordingly [33].

Validation With Treatments of Reference

We compared the data-driven determination of adjustment sets with a knowledge-based approach on 3 treatments of reference: first, dexamethasone, for which we expected a beneficial effect on 28-day mortality and which we assumed is subject to strong indication bias, and second and third, drugs of reference with an expected null effect, with high prevalence (paracetamol) and low prevalence (phloroglucinol). We studied the association of these treatments of reference with 28-day mortality on the overall population and in age-based subgroups (patients <70 or ≥70 years old). Indeed, age is the most important prognosis factor in COVID-19, and dexamethasone's beneficial effect is heterogeneous across age subgroups [3].

We compared the association after adjusting on the data-driven adjustment set. For the knowledge-based approach, we used a set of known prognosis factors extracted from Medline articles, including age, gender, number of comorbidities, platelet count, prothrombin ratio, creatinine, blood urea nitrogen, C-reactive protein (CRP), mean arterial pressure, systolic arterial pressure, and peripheral capillary oxygen saturation [34-37].

Missing Data Management Strategy

Missing data management followed a 2-step strategy targeting 2 different missing data mechanisms. In the first step, we excluded patients with a number of observations lower than the 2nd percentile for drugs, laboratory tests, or physiological measurements. A comparison of baseline characteristics between patients included in the analysis and patients excluded for missing data was performed to detect a possible selection bias. In the second step, we performed multiple imputation with chained equations (MICE) [38]. MICE was performed using 5 imputed data sets, and the predictive mean matching strategy was chosen, using all adjustment covariates (ie, not including drugs). The adjustment set selection was adapted to the setting of multiple imputed data sets by selecting variables that appeared in at least half of the imputed data sets. ORs were pooled according to the Rubin rule after log transformation.

In addition to these missing data–handling strategies, we also reported a measure of data missingness specific to each model, the fraction of missing information (FMI), which is considered moderately large above 0.3 and high above 0.5 [39].

Implementation

Analyses were performed using R statistical software version 3.5.1 (R Core Team) [40]. The following packages were combined in custom functions to provide a reproducible and configurable pipeline: MICE [41], glmnet [42], MatchIt [43], and PSWeight [44]. The code is available online for transparency [45].

Ethics

This study was approved by the Institutional Review Board of the AP-HP CDW (reference CSE-20-18-COVID19). All patients admitted to the AP-HP were informed of the possible reuse of their EHRs for research purposes according to the European General Data Protection Regulation and had the right to opt out of participating, in agreement with the Commission Nationale

de l'Informatique et des Libertés (regulatory decision DE-2018-155).

Results

Population Characteristics

Of 39 different hospitals, 16 (41%) were retained for the study based on the availability of drug exposure information from computerized physician order entries. In these 16 hospitals, we found a total of 8922 eligible patients, of which 3139 (35.18%) were excluded because of insufficient information regarding drug exposure, lab tests, or physiological measurements (see Figure 1). Included and excluded patients were comparable for age (median age 69.2 [IQR 56.7-81.1] vs 70.9 [IQR 55.8-83.8]

years) and number of comorbidities (2731/5783 [47.22%] vs 1591/3139 [50.68%] patients with at least 3 comorbidities) but were more often male (3390/5783 [58.62%] vs 1599/3139 [50.94%]); see Table S1 in Multimedia Appendix 1.

A total of 5783 patients were included in the analysis with a median age at admission of 69.2 (IQR 56.7-81.1) years, and 3390 (58.62%) of them were male (Table 1). Patients were admitted from 16 hospitals, with 3 (19%) hospitals representing 2758 (47.69%) of patients. Frequent comorbidities included hypertension (n=2065, 35.71%), chronic kidney disease (n=554, 9.58%), atrial fibrillation or flutter (n=458, 7.92%), dyslipidemia (n=357, 6.17%), and ischemic chronic heart disease (n=356, 6.16%); see Table 1.

Table 1. Baseline characteristics of the population (N=5783).

Characteristics	Value
Age at diagnostic (years), median (Q1, Q3)	69.2 (56.7, 81.1)
Age group at diagnostic (years), n (%)	
18-39	322 (5.57)
40-49	538 (9.30)
50-59	948 (16.39)
60-69	1184 (20.47)
70-79	1241 (21.46)
80+	1550 (26.80)
Gender (male), n (%)	3390 (58.62)
Deaths, n (%)	933 (16.13)
Follow-up (days), median (Q1, Q3)	8.8 (5.2, 14.9)
28-day Mortality, n (%)	635 (10.98)
Center, n (%)	
GH A Chenevier-H Mondor	965 (16.69)
Hôpital Saint Antoine	887 (15.34)
Hôpital Tenon	849 (14.68)
Other	3082 (53.29)
Time period, n (%)	
February-July 2020	2187 (37.82)
August-November 2020	1197 (20.70)
December 2020-June 2021	2399 (41.48)
Comorbidities, n (%)	
Hypertension	2065 (35.71)
Severe protein energy malnutrition	655 (11.33)
Chronic kidney disease	554 (9.58)
Light or moderate protein energy malnutrition	509 (8.80)
Atrial fibrillation and flutter	458 (7.92)
Dyslipidemia	357 (6.17)
Ischemic chronic heart disease	356 (6.16)
Deficiency in vitamin D	339 (5.86)
Presence of cardiac and vascular implants and grafts	306 (5.29)
Hypothyroidism, unspecified	288 (4.98)
Other parameters, median (Q1, Q3)	
BMI	26.5 (23.4, 30.3)
Pulsations (/min)	89 (78, 102)
Diastolic arterial pressure (mmHg)	76 (66, 85)
Systolic arterial pressure (mmHg)	131 (117, 146)
Respiratory rate (/min)	24 (20, 28)
Peripheral capillary oxygen saturation (%)	95 (92, 97)
Body temperature (°C)	37.4 (36.8, 38.2)
Hemoglobin (g/dL)	13.10 (11.80, 14.30)
White blood cell count (10^9/L)	6.38 (4.79, 8.51)

Characteristics	Value
Creatinine (μmol/L)	80 (65.00, 105.50)
Blood urea nitrogen (mmol/L)	6.40 (4.60, 9.50)
CRP[a] (mg/L)	69.80 (32.50, 122.20)
Oxygen blood saturation (%)	95 (92.70, 97.00)
Fibrinogen (g/L)	5.80 (4.85, 6.82)
Bicarbonate (mmol/L)	25 (22.00, 27.60)

[a]CRP: C-reactive protein.

Validation With Treatment of References

Without adjustment, dexamethasone was associated with decreased 28-day mortality in patients under 70 years old (OR 0.40, 95% CI 0.26-0.62, $P<.001$) and with increased 28-day mortality in patients over 70 years old (OR 1.40, 95% CI 1.14-1.71, $P<.001$).

Adjusting using the "known PF" and "data driven" adjustments sets yielded close results, except for dexamethasone in patients over 70 years old. In the latter subgroup, only the "data driven" adjustment set yielded no association of dexamethasone with increased 28-day mortality (Figure 2).

On the matched subset for dexamethasone in patients under 70 years old, 3158 (54.61%) of 5783 exposed patients found matches, the association with mortality was strong (OR 0.41, 95% CI 0.21-0.79, $P=.01$), but the FBC was only 35%. On the "weighted and trimmed" subset, the FEP that fell in the EER was only 23.1%, the association with mortality was no longer significant (OR 0.46, 95% CI 0.18-1.18, $P=.1$), but 100% of covariates were balanced.

The fraction of missing information was low and never exceeded 0.2. The FEP was lower than 50% for dexamethasone in all subgroups.

Figure 2. Treatment of references for validating data-driven adjustment set selection. The association between 28-day mortality and early exposure to treatment was measured as the ORs for 3 treatments of reference: (1) dexamethasone with expected beneficial effect on 28-day mortality and (2) treatments with an expected null effect, with high prevalence (paracetamol) or low prevalence (phloroglucinol). We compared 2 pretreatment covariate sets: "known PF" using PFs from the literature (blue) and "data driven" for a model selection procedure based on adaptive LASSO (green) targeting confusion factors. ORs were computed by logistic regression on the overall data set (red), matched or weighted and trimmed subpopulations based on PSs. LASSO: least absolute shrinkage and selection operator; OR: odds ratio; PF: prognostic factor; PS: propensity score. *$P<.05$; **$P<.01$; ***$P<.001$.
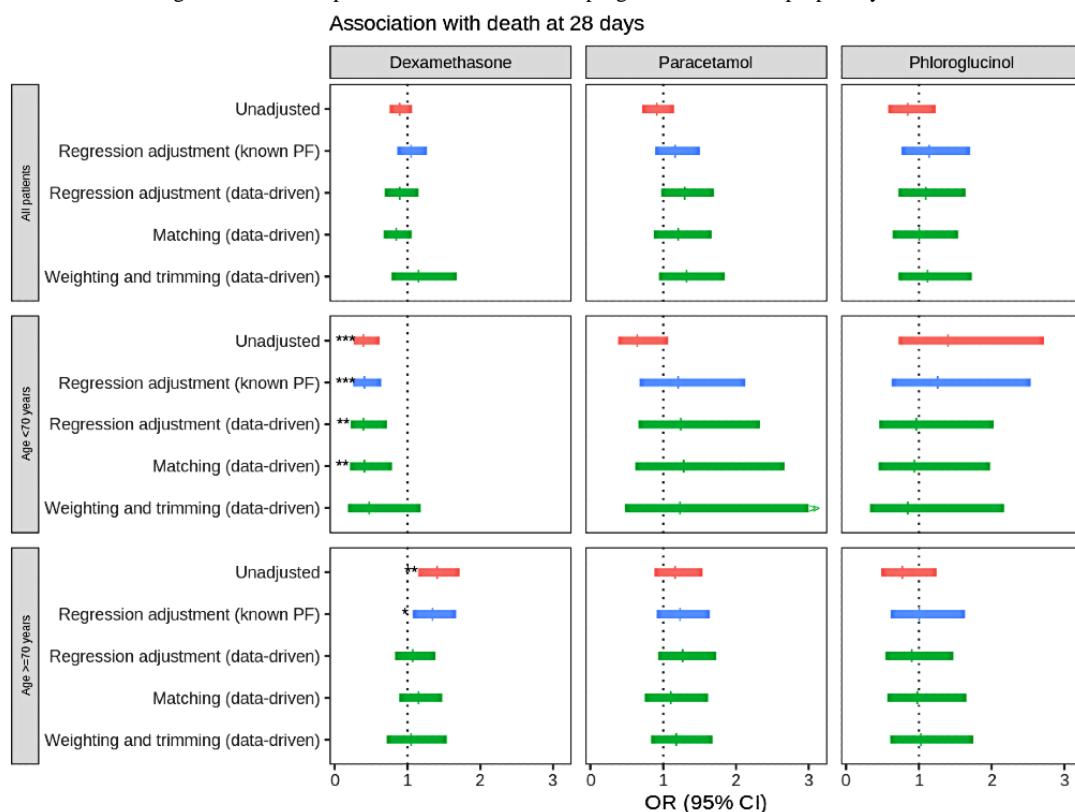
Pharmacopeia-wide Association With 28-day Mortality

Primary Analysis

We identified a total of 87 different drugs (ATC level 5codes, eg, B01AF01 rivaroxaban) administered within the first 48 hours and present in at least 100 patient records (Figure 3). Detailed results are given in Figure 4 for drugs with *P*<.15. After correction for multiple hypothesis testing, none were associated with reduced in-hospital mortality, and 4 (5%) remained associated with increased in-hospital mortality on the overall population after adjustment: sulfamethoxazole-trimethoprim, valaciclovir, tramadol, and diazepam (Table 2). Analyses of matched subpopulations found consistent results, with a good fraction of covariate balance (between 98% and 100% of covariates with ASMD<0.1), except for diazepam (89%). Analyses of weighted subpopulations were not consistent and found a small FEP for sulfamethoxazole-trimethoprim and valaciclovir.

Figure 3. Pharmacopeia-wide association with 28-day mortality. Each dot represents the FDR-corrected *P* value (*Q* value), on a negative log scale (y axis) of a drug (ATC code), on the x axis. An ATC code is attributed if the drug is prescribed in the first 48 hours of COVID-19 admission in conventional wards. The color indicates the pharmacological subgroup (ATC level 2). The top panel reports *Q* values from the primary analysis, using a multivariate logistic regression model, and the dotted line indicates a 5% FDR. The middle and bottom panels report secondary analyses using matching and inverse probability weighting methods, respectively. Dot sizes are inversely proportional to *Q* values. ATC: anatomical therapeutic chemical; FDR: false discovery rate.
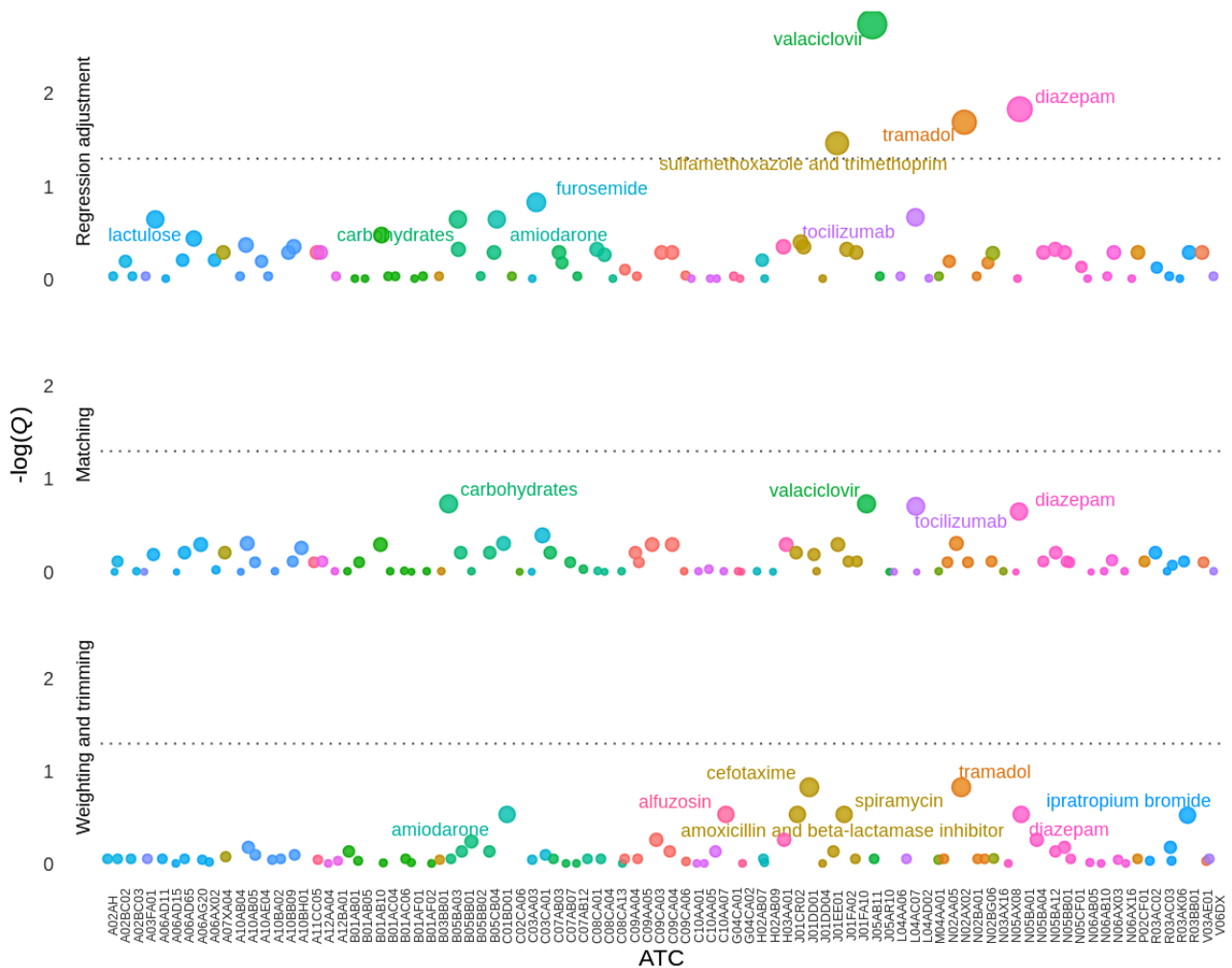
Figure 4. Increased and decreased mortality for the top drugs. Association is reported as the OR between treatment exposure and 28-day mortality in different settings: without adjustment, after adjusting, and on matched and weighted subpopulations based on treatment-specific PSs. p-values are indicated without multiple hypothesis testing correction. Treatments are ordered from top to bottom by decreasing adjusted OR. Drugs at the top tend to be associated with increased mortality, while drugs at the bottom tend to be associated with decreased mortality. Colors correspond to ATC level 2. Only drugs with *P*<.15 are reported. ATC: anatomical therapeutic chemical; OR: odds ratio; PS: propensity score. *P<.05; **P<.01; ***P<.001.
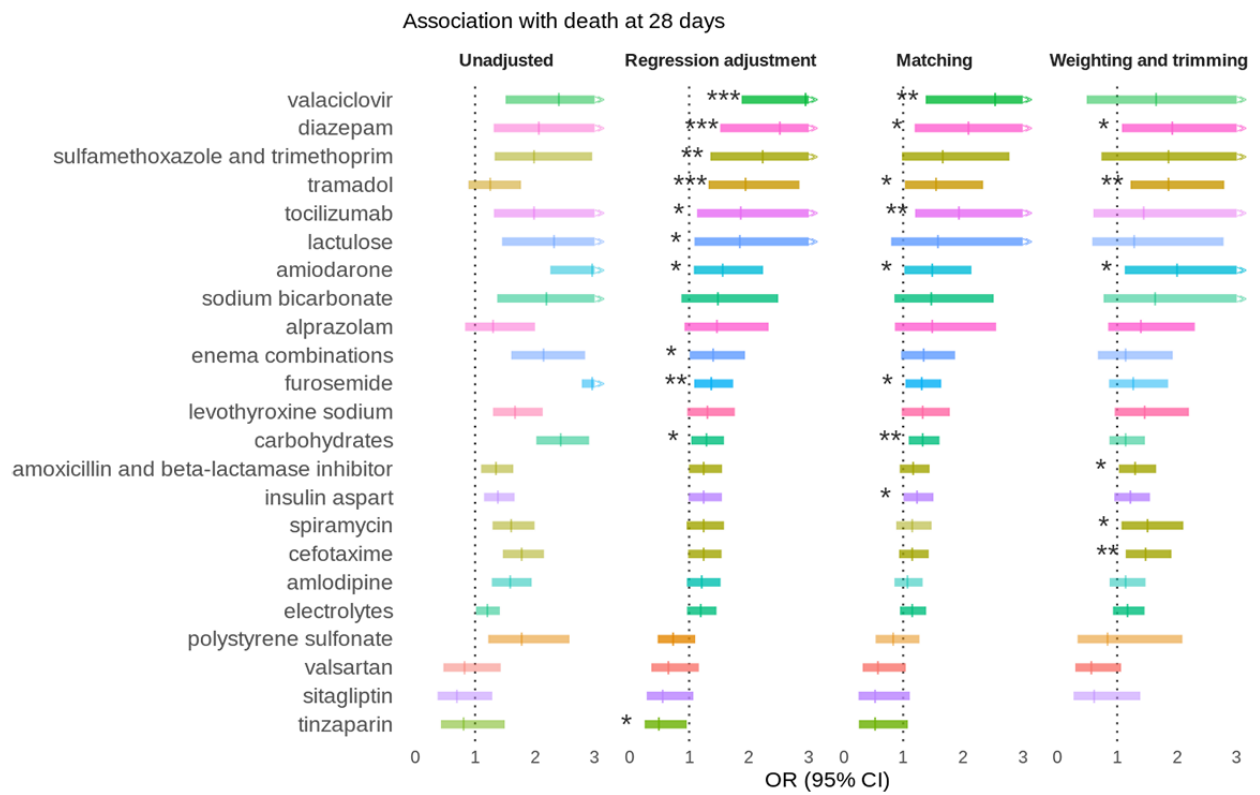
Table 2. Treatment associated with 28-day mortality after regression adjustment at 5% FDR[a].

Tests	Treated vs controls (events/exposed), n/n	OR[b] (95% CI)	Q value[c]	FBC[d] (%)	FEP[e] (%)	FMI[f]
Sulfamethoxazole and trimethoprim						
Regression adjustment	31/161 vs 604/5622	2.22 (1.36-3.64)	.03	N/A[g]	100	<0.01
Matching	31/160 vs 63/518	1.65 (0.98-2.78)	N/A	99	99.3	0.03
Weighting and trimming	8/51 vs 172/1790	1.86 (0.74-4.68)	N/A	91	32.1	0.02
Valaciclovir						
Regression adjustment	24/107 vs 611/5676	3.21 (1.88-5.48)	.002	N/A	100	0.01
Matching	24/107 vs 41/404	2.54 (1.38-4.67)	N/A	98	99.4	0.08
Weighting and trimming	6/35 vs 210/2136	1.64 (0.49-5.51)	N/A	71	32.5	0.16
Tramadol						
Regression adjustment	40/302 vs 595/5481	1.94 (1.32-2.85)	.02	N/A	100	<0.01
Matching	40/301 vs 108/1191	1.55 (1.03-2.34)	N/A	100	99.7	0.08
Weighting and trimming	31/223 vs 362/4002	1.85 (1.22-2.79)	N/A	100	74	0.02
Diazepam						
Regression adjustment	24/120 vs 611/5663	2.51 (1.52-4.16)	.01	N/A	100	<0.01
Matching	23/116 vs 47/448	2.09 (1.19-3.66)	N/A	89	96.7	0.06
Weighting and trimming	15/89 vs 483/4925	1.92 (1.08-3.41)	N/A	100	74.3	0.01

[a]FDR: false discovery rate.

[b]OR: odds ratio

[c]*Q* value: FDR-corrected *P* value.

[d]FBC: fraction of balanced covariates.

[e]FEP: fraction of exposed population.

[f]FMI: fraction of missing information.

[g]N/A: not applicable.

Secondary Analysis

We highlight here the results of the weighted and trimmed population where patients were more comparable (Figure 4). Interestingly, 2 angiotensin receptor blockers (ARBs) came up as the top 5 treatments with OR<1, with treatments ordered by *P* values (Table 3). We further explored this hypothesis and found that in weighted and trimmed analysis, ARBs with a high affinity for angiotensin receptor 1 (dissociation constant $K_d \geq 6$: telmisartan, valsartan, losartan) tended to be associated with decreased 28-day mortality compared to ARBs with a lower affinity ($K_d < 6$: irbesartan, candesartan, olmesartan)—OR 0.56 (95% CI 0.34-0.91).

Table 3. Top 5 treatments with OR[a]<1 in the weighted and trimmed population, ordered by *P* value. None were significantly associated with mortality after FDR[b] correction in the primary analysis.

Treatment	Treated vs controls (events/exposed), n/n	OR (95% CI)	FBC[c] (%)	FEP[d] (%)	FMI[e]
Sitagliptin	7/100 vs 426/4023	0.61 (0.27-1.39)	100	71.7	0.01
Valsartan	11/132 vs 562/4567	0.57 (0.30-1.06)	100	87.1	0.01
Irbesartan	32/277 vs 538/4527	0.77 (0.52-1.13)	100	91.4	0.01
Rosuvastatin	12/131 vs 399/3214	0.64 (0.35-1.19)	100	60	0.01
Alfuzosin	6/100 vs 182/1875	0.34 (0.13-0.84)	100	39.3	0.03

[a]OR: odds ratio.

[b]FDR: false discovery rate.

[c]FBC: fraction of balanced covariates.

[d]FEP: fraction of exposed population.

[e]FMI: fraction of missing information.

Discussion

Principal Findings

We systematically assessed the association of early in-hospital treatments with 28-day mortality in a large, multicenter retrospective case study of 5783 patients with COVID-19, using an innovative PheWAS-like approach. We showed empirical evidence that our fully data-driven pipeline is comparable to or better than a knowledge-based approach to adjust for confounding on 3 drugs of reference. We showed in this practical implementation for COVID-19 how such a pipeline can be used to mine EHR pharmacopoeia and generate pharmacological hypotheses in an exploratory fashion. Indeed, of 87 treatments prescribed in the first 48 hours, 4 (5%) were associated with increased 28-day mortality after adjustment of confounding factors and multiple testing correction, and none were associated with decreased mortality. Among those 4, only diazepam and tramadol had consistent results in secondary analyses more robust to confounding. In addition, secondary analyses suggested that high-affinity ARBs are associated with reduced COVID-19 28-day mortality, suggesting they may be beneficial for patients with COVID-19.

Validation With Drugs of Reference

We tested our adjustment methods on treatments for which the effect on 28-day mortality is documented (protective effect for dexamethasone) or unlikely to be different from null (absence of an effect for paracetamol and phloroglucinol). Subgroup analysis of the Randomised Evaluation of Covid-19 Therapy (RECOVERY) trial suggests that patients under 70 years old only benefit from dexamethasone, with OR 0.64 (95% CI 0.53–0.78) versus OR 1.03 (95% CI 0.84–1.25) between 70 and 80 years old and OR 0.89 (95% CI 0.75–1.05) above 80 years old [3]. Our automated pipeline finds overall consistent results between the "data driven" and the "known PF" adjustment set for the 3 drugs of reference. The differences observed for the dexamethasone effect on the >70-year age group could be explained by missing or misspecified confounding factors in the "known PF" adjustment set compared to the "data driven" adjustment set (see Table S2 in Multimedia Appendix 1). Overall, these results provide empirical evidence that the automated determination of adjustment sets on these 3 drugs yields valid adjustment sets, sufficient for controlling indication biases.

Pharmacopeia-wide Association With 28-day Mortality

Interestingly, diazepam, an anxiolytic benzodiazepine, was found to be associated with a detrimental effect on in-hospital mortality in our study. This result might not be COVID-19 specific, as benzodiazepines have shown a dose-response association with mortality in patients with severe chronic obstructive pulmonary disease [46]. We also found that tramadol, a weak opioid, is associated with increased 28-day mortality. Noteworthy, both benzodiazepines and tramadol may have adverse respiratory effects, such as respiratory depression, which, although not specific to COVID-19, could be detrimental in patients with severe COVID-19 pneumonitis [47]. In addition, our automated pipeline allowed us to generate a pharmacological hypothesis consistent with results from a clinical trial. Indeed,

an open randomized controlled trial showed that death by day 30 was reduced in patients undergoing telmisartan therapy (control: 16/71 [23%]; telmisartan: 3/70 [4%] participants; *P*=0.002) hospitalized for COVID-19 [48]. However, other studies did not find an association between ARBs and COVID-19 mortality, and further studies are needed to assess this finding and investigate potential mechanisms [49].

Limits and Strengths

This retrospective study methodology was based on reusing routinely collected clinical data across 16 hospitals of the Greater Paris area. Unexpectedly, from an initial set of 8922 COVID-19 patient records, only 5783 (64.82%) patient records ended up meeting all inclusion criteria. However, this is not intrinsically linked to our method but rather to the relative lack of maturity of the hospitals' information systems, particularly concerning drug prescription. Indeed, at the beginning of the pandemic, computerized physician order entry was not available in all hospitals and units or not linked to the CDW. Although this study followed most of the guidelines provided by Kohane et al [50], such as a multidisciplinary approach, code transparency, and robustness against variability across hospitals, this result stresses that data completeness in EHRs remains an open question. We can hypothesize that the pandemic will have a boosting effect on the maturation of the information system of hospitals. Regarding confusion adjustment, we could have used more flexible models to fit PSs, such as random forests, and used double robust estimators, which are less sensible to model misspecification [26]. However, we found that the most important factors for accurate measures of treatment association with mortality were the choice of adjustment sets and the use of trimming. Moreover, we decided to restrict to methods that would easily scale to large sets of exposures. Globally, our results are dependent, as in all complex analysis of real-life data, on choices in the preprocessing and modelling of the data. These dependencies can be subtle and lead to changes in amplitude or direction of the measured associations, sometimes framed as "vibration of effect" [51]. Our rationale was to decide these questions based on theoretical grounds (or simulation studies) to leverage treatment of references if not possible (eg, data driven vs knowledge based) and finally to report multiple analyses if uncertainty remains about which method is more relevant (eg, matching or inverse probability weighting).

Large-scale association studies such as this work are known to require a large amount of data to reveal significant associations. Therefore, it may be difficult to obtain sufficient statistical power. To get around this difficulty, it is possible to run the association test using aggregated data to an upper level in the ATC.

Regarding clinical significance, COVID-19 is a biphasic disease, with a viral replication period and then an inflammatory state, and patients may not be hospitalized at the same time of the disease. This may have led to heterogeneity in the condition of the patients and complicated the interpretation of the results. Furthermore, there is a potential risk of selection bias since we dropped 35% of COVID-19 admissions due to data missingness. However, excluded patients were comparable in terms of age and number of comorbidities to the patients included in this

study (Table S1 in Multimedia Appendix 1), which is in favor of the generalizability of the obtained results. In addition, we cannot rule out that some confounding factors remain unobserved. Secondary analyses based on EER allowed us to partially address this issue, since the sample size remained large enough in this setting to include a broad amount of potential confounding, and we analyzed a rather homogeneous population by excluding patients admitted to intensive care units (ICUs) in the first 48 hours.

The main strength of our study lies in its external validity: it used data collected across 16 different hospitals of the Greater Paris area and included a large number of patients with COVID-19. These characteristics make it likely to capture the variability of populations and disease management in real-life settings. Similarly, we addressed treatment-specific indication biases in a fully data-driven fashion, which we validated empirically on drugs of reference. This methodology based on a hypothesis-free exploration of COVID-19-related EHRs is easily exportable to other settings. Population trimming based on stabilized PSs allowed us to restrict the analysis to comparable patients, which cannot be done by a simple outcome-oriented regression adjustment. Finally, it allowed us to generate a measure of covariate balancing, which turns out to be a critical diagnostic for studying a large array of drug-outcome associations.

Our systematic hypothesis-free approach constitutes a promising tool that can be rapidly used in the setting of emergent diseases to generate potential drug candidates. Still, these drug candidates need to be further assessed from a pharmacological point of view before being tested in clinical trials. Further developments will include time dependency of treatments, covariates, and outcomes in a more flexible way, not restricted to landmark analysis (28-day mortality) and window-type restriction of exposition. In addition, including information from the natural language processing (NLP) extraction workflow would largely enrich such a pipeline [50,52].

Conclusion

Our innovative approach proved useful in rapidly generating pharmacological hypotheses in an outbreak setting, without requiring a priori knowledge of the disease. Our systematic analysis of early prescribed treatments from patients hospitalized for COVID-19 showed that diazepam and tramadol are associated with increased 28-day mortality. Whether these drugs could worsen COVID-19 needs to be further assessed.

Data Availability

The aggregated results data will be published online, along with the code generating it [53].

Authors' Contributions

IL contributed to conceptualization, data curation, formal analysis, methodology, software, and writing (original draft). ASL contributed to methodology and writing (reviewing and editing). LC contributed to conceptualization, formal analysis, and writing (reviewing and editing). BR and NG contributed to validation, and writing (reviewing and editing). AB contributed to project administration, resources, supervision, validation, and writing (reviewing and editing). AN contributed to conceptualization, data curation, formal analysis, software, resources, supervision, validation, writing (original draft), and writing (reviewing and editing).

Conflicts of Interest

None declared.

Multimedia Appendix 1

XSL•FO

RenderX

Supplementary materials.
[DOCX File , 112 KB - medinform_v10i3e35190_app1.docx]

References

1. COVID-19 Dashboard by the Center for Systems Science and Engineering (CSSE) at Johns Hopkins University (JHU). URL: https://coronavirus.jhu.edu/map.html [accessed 2021-08-27]

2. Boutron I, Chaimani A, Devane D, Meerpohl J, Tovey D, Hróbjartsson A, et al. Interventions for preventing and treating COVID-19: protocol for a living mapping of research and a living systematic review. Cochrane Database Syst Rev 2020(11):Art. No.: CD013769. [doi: 10.1002/14651858.CD013769]

3. RECOVERY Collaborative Group. Dexamethasone in hospitalized patients with Covid-19. N Engl J Med 2021 Feb 25;384(8):693-704. [doi: 10.1056/nejmoa2021436]

4. Gérard A, Romani S, Fresse A, Viard D, Parassol N, Granvuillemin A, French Network of Pharmacovigilance Centers. "Off-label" use of hydroxychloroquine, azithromycin, lopinavir-ritonavir and chloroquine in COVID-19: a survey of cardiac adverse drug reactions by the French Network of Pharmacovigilance Centers. Therapie 2020 Jul;75(4):371-379 [FREE Full text] [doi: 10.1016/j.therap.2020.05.002] [Medline: 32418730]

5. Wong AY, MacKenna B, Morton CE, Schultze A, Walker AJ, Bhaskaran K, OpenSAFELY Collaborative. Use of non-steroidal anti-inflammatory drugs and risk of death from COVID-19: an OpenSAFELY cohort analysis based on two cohorts. Ann Rheum Dis 2021 Jul 21;80(7):943-951 [FREE Full text] [doi: 10.1136/annrheumdis-2020-219517] [Medline: 33478953]

6. Chouchana L, Beeker N, Garcelon N, Rance B, Paris N, Salamanca E, AP-HP/Universities/Inserm COVID-19 research collaboration, AP-HP Covid CDR Initiative,"Entrepôt de Données de Santé" AP-HP Consortium". Association of antihypertensive agents with the risk of in-hospital death in patients with Covid-19. Cardiovasc Drugs Ther 2021 Feb 17;17:1-6 [FREE Full text] [doi: 10.1007/s10557-021-07155-5] [Medline: 33595761]

7. Hellwig MD, Maia A. A COVID-19 prophylaxis? Lower incidence associated with prophylactic administration of ivermectin. Int J Antimicrob Agents 2021 Jan;57(1):106248 [FREE Full text] [doi: 10.1016/j.ijantimicag.2020.106248] [Medline: 33259913]

8. Caly L, Druce JD, Catton MG, Jans DA, Wagstaff KM. The FDA-approved drug ivermectin inhibits the replication of SARS-CoV-2 in vitro. Antiviral Res 2020 Jun;178:104787 [FREE Full text] [doi: 10.1016/j.antiviral.2020.104787] [Medline: 32251768]

9. Wang M, Cao R, Zhang L, Yang X, Liu J, Xu M, et al. Remdesivir and chloroquine effectively inhibit the recently emerged novel coronavirus (2019-nCoV) in vitro. Cell Res 2020 Mar 04;30(3):269-271 [FREE Full text] [doi: 10.1038/s41422-020-0282-0] [Medline: 32020029]

10. Siuka D, Pfeifer M, Pinter B. Vitamin D supplementation during the COVID-19 pandemic. Mayo Clin Proc 2020 Aug;95(8):1804-1805 [FREE Full text] [doi: 10.1016/j.mayocp.2020.05.036] [Medline: 32753156]

11. Nopp S, Moik F, Jilma B, Pabinger I, Ay C. Risk of venous thromboembolism in patients with COVID-19: a systematic review and meta-analysis. Res Pract Thromb Haemost 2020 Sep 25;4(7):1178-1191 [FREE Full text] [doi: 10.1002/rth2.12439] [Medline: 33043231]

12. Cohen AT, Davidson BL, Gallus AS, Lassen MR, Prins MH, Tomkowski W, et al. Efficacy and safety of fondaparinux for the prevention of venous thromboembolism in older acute medical patients: randomised placebo controlled trial. BMJ 2006 Jan 26;332(7537):325-329. [doi: 10.1136/bmj.38733.466748.7c]

13. Samama MM, Cohen AT, Darmon J, Desjardins L, Eldor A, Janbon C, et al. A comparison of enoxaparin with placebo for the prevention of venous thromboembolism in acutely ill medical patients. N Engl J Med 1999 Sep 09;341(11):793-800 [FREE Full text] [doi: 10.1056/nejm199909093411103]

14. Denny JC, Ritchie MD, Basford MA, Pulley JM, Bastarache L, Brown-Gentry K, et al. PheWAS: demonstrating the feasibility of a phenome-wide scan to discover gene-disease associations. Bioinformatics 2010 May 01;26(9):1205-1210 [FREE Full text] [doi: 10.1093/bioinformatics/btq126] [Medline: 20335276]

15. Denny J, Bastarache L, Ritchie M, Carroll R, Zink R, Mosley J. Systematic comparison of phenome-wide association study of electronic medical record data and genome-wide association study data. Nat Biotechnol 2013;31:1110. [doi: 10.3410/f.718185859.793488460]

16. Neuraz A, Chouchana L, Malamut G, Le Beller C, Roche D, Beaune P, et al. Phenome-wide association studies on a quantitative trait: application to TPMT enzyme activity and thiopurine therapy in pharmacogenomics. PLoS Comput Biol 2013 Dec 26;9(12):e1003405 [FREE Full text] [doi: 10.1371/journal.pcbi.1003405] [Medline: 24385893]

17. Ryan P, Madigan D, Stang P, Schuemie M, Hripcsak G. Medication-wide association studies. CPT Pharmacomet Syst Pharmacol 2013 Sep 18;2(9):e76 [FREE Full text] [doi: 10.1038/psp.2013.52] [Medline: 24448022]

18. Patel CJ, Ji J, Sundquist J, Ioannidis JPA, Sundquist K. Systematic assessment of pharmaceutical prescriptions in association with cancer risk: a method to conduct a population-wide medication-wide longitudinal study. Sci Rep 2016 Aug 10;6(1):31308 [FREE Full text] [doi: 10.1038/srep31308] [Medline: 27507038]

XSL•FO
RenderX

19.   Choi L, Carroll R, Beck C, Mosley J, Roden D, Denny J, et al. Evaluating statistical approaches to leverage large clinical
       datasets for uncovering therapeutic and adverse medication effects. Bioinformatics 2018 Sep 01;34(17):2988-2996 [FREE
       Full text] [doi: 10.1093/bioinformatics/bty306] [Medline: 29912272]
20.   Bejan CA, Cahill KN, Staso PJ, Choi L, Peterson JF, Phillips EJ. DrugWAS: drug-wide association studies for COVID-19
       drug repurposing. Clin Pharmacol Ther 2021 Dec 10;110(6):1537-1546 [FREE Full text] [doi: 10.1002/cpt.2376] [Medline:
       34314511]
21.   Austin PC, Grootendorst P, Anderson GM. A comparison of the ability of different propensity score models to balance
       measured variables between treated and untreated subjects: a Monte Carlo study. Stat Med 2007 Feb 20;26(4):734-753.
       [doi: 10.1002/sim.2580] [Medline: 16708349]
22.   Brookhart M, Schneeweiss S, Rothman KJ, Glynn RJ, Avorn J, Stürmer T. Variable selection for propensity score models.
       Am J Epidemiol 2006 Jun 15;163(12):1149-1156 [FREE Full text] [doi: 10.1093/aje/kwj149] [Medline: 16624967]
23.   Witte J, Didelez V. Covariate selection strategies for causal inference: classification and comparison. Biom J 2019 Sep
       10;61(5):1270-1289. [doi: 10.1002/bimj.201700294] [Medline: 30306605]
24.   Stürmer T, Rothman KJ, Avorn J, Glynn RJ. Treatment effects in the presence of unmeasured confounding: dealing with
       observations in the tails of the propensity score distribution; a simulation study. Am J Epidemiol 2010 Oct 01;172(7):843-854
       [FREE Full text] [doi: 10.1093/aje/kwq198] [Medline: 20716704]
25.   Stürmer T, Webster-Clark M, Lund J, Wyss R, Ellis A, Lunt M, et al. Propensity score weighting and trimming strategies
       for reducing variance and bias of treatment effect estimates: a simulation study. Am J Epidemiol 2021 Aug
       01;190(8):1659-1670 [FREE Full text] [doi: 10.1093/aje/kwab041] [Medline: 33615349]
26.   Rothman KJ, Lanes S, Robins J. Causal inference. Epidemiology 1993;4(6):555. [doi: 10.1097/00001648-199311000-00013]
27.   Walker A, Patrick A, Lauer M, Hornbrook M, Marin M, Platt R, et al. A tool for assessing the feasibility of comparative
       effectiveness research. CER 2013 Jan:11 [FREE Full text] [doi: 10.2147/cer.s40357]
28.   Daniel C, Serre P, Orlova N, Bréant S, Paris N, Griffon N. Initializing a hospital-wide data quality program. The AP-HP
       experience. Comput Methods Programs Biomed 2019 Nov;181:104804. [doi: 10.1016/j.cmpb.2018.10.016] [Medline:
       30497872]
29.   Anatomical Therapeutic Chemical (ATC) Classification. URL: https://www.who.int/tools/atc-ddd-toolkit/atc-classification
       [accessed 2021-05-02]
30.   Zou H. The adaptive Lasso and its oracle properties. J Am Stat Assoc 2012 Jan 01;101(476):1418-1429 [FREE Full text]
       [doi: 10.1198/016214506000000735]
31.   Austin PC. Optimal caliper widths for propensity-score matching when estimating differences in means and differences in
       proportions in observational studies. Pharm Stat 2011 Mar 29;10(2):150-161 [FREE Full text] [doi: 10.1002/pst.433]
       [Medline: 20925139]
32.   Stuart EA, Lee BK, Leacy FP. Prognostic score-based balance measures can be a useful diagnostic for propensity score
       methods in comparative effectiveness research. J Clin Epidemiol 2013 Aug;66(8 Suppl):S84-S90.e1 [FREE Full text] [doi:
       10.1016/j.jclinepi.2013.01.013] [Medline: 23849158]
33.   Benjamini Y, Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. J R
       Stat Soc: Series B (Methodol) 2018 Dec 05;57(1):289-300. [doi: 10.1111/j.2517-6161.1995.tb02031.x]
34.   Zheng Z, Peng F, Xu B, Zhao J, Liu H, Peng J, et al. J Infect 2020 Aug;81(2):e16-e25 [FREE Full text] [doi:
       10.1016/j.jinf.2020.04.021] [Medline: 32335169]
35.   Williamson EJ, Walker AJ, Bhaskaran K, Bacon S, Bates C, Morton CE, et al. Factors associated with COVID-19-related
       death using OpenSAFELY. Nature 2020 Aug 08;584(7821):430-436 [FREE Full text] [doi: 10.1038/s41586-020-2521-4]
       [Medline: 32640463]
36.   Gue YX, Tennyson M, Gao J, Ren S, Kanji R, Gorog DA. Development of a novel risk score to predict mortality in patients
       admitted to hospital with COVID-19. Sci Rep 2020 Dec 07;10(1):21379 [FREE Full text] [doi: 10.1038/s41598-020-78505-w]
       [Medline: 33288840]
37.   Di Castelnuovo A, Bonaccio M, Costanzo S, Gialluisi A, Antinori A, Berselli N, COvid-19 RISkTreatments (CORIST)
       collaboration. Common cardiovascular risk factors and in-hospital mortality in 3,894 patients with COVID-19: survival
       analysis and machine learning-based findings from the multicentre Italian CORIST Study. Nutr Metab Cardiovasc Dis
       2020 Oct 30;30(11):1899-1913 [FREE Full text] [doi: 10.1016/j.numecd.2020.07.031] [Medline: 32912793]
38.   van Buuren S. Flexible Imputation of Missing Data, Second Edition. Boca Raton, FL: Chapman and Hall/CRC; 2018.
39.   Li KH, Raghunathan TE, Rubin DB. Large-sample significance levels from multiply imputed data using moment-based
       statistics and an F reference distribution. J Am Stat Assoc 1991 Dec;86(416):1065 [FREE Full text] [doi: 10.2307/2290525]
40.   R Core Team. R: A Language and Environment for Statistical Computing. Vienna, Austria: R Foundation for Statistical
       Computing; 2018.
41.   Buuren SV, Groothuis-Oudshoorn K. mice: Multivariate imputation by chained equations. J Stat Softw 2011;45(3):20.
       [doi: 10.18637/jss.v045.i03]
42.   Friedman J, Hastie T, Tibshirani R. Regularization paths for generalized linear models via coordinate descent. J Stat Softw
       2010;33(1):1-22. [doi: 10.18637/jss.v033.i01]

43. Ho DE, Imai K, King G, Stuart EA. MatchIt: nonparametric preprocessing for parametric causal inference. J Stat Softw 2011;42(8):7. [doi: 10.18637/jss.v042.i08]

44. Zhou T, Tong G, Li F, Thomas L. PSweight: An R Package for Propensity Score Weighting Analysis. 2020. URL: http://arxiv.org/abs/2010.08893 [accessed 2021-08-27]

45. pharmwas. URL: https://gitlab.com/lerner.ivan/pharmwas [accessed 2021-11-29]

46. Ekström MP, Bornefalk-Hermansson A, Abernethy AP, Currow DC. Safety of benzodiazepines and opioids in very severe respiratory disease: national prospective study. BMJ 2014 Jan 30;348(jan30 2):g445-g445 [FREE Full text] [doi: 10.1136/bmj.g445] [Medline: 24482539]

47. Murray MJ, DeRuyter ML, Harrison BA. Opioids and benzodiazepines. Crit Care Clin 1995 Oct;11(4):849-873 [FREE Full text] [doi: 10.1016/s0749-0704(18)30042-3]

48. Duarte M, Pelorosso F, Nicolosi LN, Salgado MV, Vetulli H, Aquieri A, et al. Telmisartan for treatment of Covid-19 patients: an open multicenter randomized clinical trial. EClinicalMedicine 2021 Jul;37:100962 [FREE Full text] [doi: 10.1016/j.eclinm.2021.100962] [Medline: 34189447]

49. Chouchana L, Beeker N, Garcelon N, Rance B, Paris N, Salamanca E. Correction to: association of antihypertensive agents with the risk of in-hospital death in patients with Covid-19. Cardiovasc Drugs Ther 2021;17:1-6 Erratum to Ref. 6. [doi: 10.1101/2020.11.23.20237362]

50. Kohane IS, Aronow BJ, Avillach P, Beaulieu-Jones BK, Bellazzi R, Bradford RL, Consortium for Clinical Characterization Of COVID-19 By EHR (4CE), et al. What every reader should know about studies using electronic health record data but may be afraid to ask. J Med Internet Res 2021 Mar 02;23(3):e22219 [FREE Full text] [doi: 10.2196/22219] [Medline: 33600347]

51. Patel CJ, Burford B, Ioannidis JP. Assessment of vibration of effects due to model specification can demonstrate the instability of observational associations. J Clin Epidemiol 2015 Sep;68(9):1046-1058 [FREE Full text] [doi: 10.1016/j.jclinepi.2015.05.029] [Medline: 26279400]

52. Neuraz A, Lerner I, Digan W, Paris N, Tsopra R, Rogier A, AP-HP/Universities/INSERM COVID-19 Research Collaboration; AP-HP COVID CDR Initiative. Natural language processing for rapid response to emergent diseases: case study of calcium channel blockers and hypertension in the COVID-19 pandemic. J Med Internet Res 2020 Aug 14;22(8):e20773 [FREE Full text] [doi: 10.2196/20773] [Medline: 32759101]

53. Lerner I. pharmwas. GitLab. URL: https://gitlab.com/lerner.ivan/pharmwas [accessed 2021-11-29]

Abbreviations

AP-HP: Assistance Publique - Hôpitaux de Paris
ARB: angiotensin receptor blocker
ASMD: absolute standardized mean difference
ATC: anatomical therapeutic chemical
CDW: clinical data warehouse
CRP: C-reactive protein
EDS: Entrepôt de Données de Santé
EER: empirical equipoise region
EHR: electronic health record
FBC: fraction of balanced covariates
FDR: false discovery rate
FEP: fraction of exposed population
FMI: fraction of missing information
ICD-10: International Classification of Diseases, 10th edition
IPW: inverse probability weighting
LASSO: least absolute shrinkage and selection operator
MICE: multiple imputation with chained equations
OR: odds ratio
PF: prognostic factor
PharmWAS: pharmacopeia-wide association study
PheWAS: phenome-wide association study
PS: propensity score
RECOVER: randomised evaluation of covid-19 therapy

<u>Original Paper</u>

The Disparity and Dynamics of Social Distancing Behaviors in Japan: Investigation of Mobile Phone Mobility Data

Zeyu Lyu[1], PhD; Hiroki Takikawa[1], PhD

Graduate School, Faculty of Arts and Letters, Tohoku University, Sendai, Japan

Corresponding Author:
Zeyu Lyu, PhD
Graduate School
Faculty of Arts and Letters
Tohoku University
27-1 Kawauchi, Aoba-ku
Sendai, 980-8576
Japan
Phone: 81 08081510072
Email: lyu.zeyu.r8@dc.tohoku.ac.jp

Abstract

Background: The availability of large-scale and fine-grained aggregated mobility data has allowed researchers to observe the dynamics of social distancing behaviors at high spatial and temporal resolutions. Despite the increasing attention paid to this research agenda, limited studies have focused on the demographic factors related to mobility, and the dynamics of social distancing behaviors have not been fully investigated.

Objective: This study aims to assist in designing and implementing public health policies by exploring how social distancing behaviors varied among various demographic groups over time.

Methods: We combined several data sources, including mobile tracking mobility data and geographical statistics, to estimate the visiting population of entertainment venues across demographic groups, which can be considered the proxy of social distancing behaviors. Next, we used time series analysis methods to investigate how voluntary and policy-induced social distancing behaviors shifted over time across demographic groups.

Results: Our findings demonstrate distinct patterns of social distancing behaviors and their dynamics across age groups. On the one hand, although entertainment venues' population comprises mainly individuals aged 20-40 years, a more significant proportion of the youth has adopted social distancing behaviors and complied with policy implementations compared to older age groups. From this perspective, the increasing contribution to infections by the youth should be more likely to be attributed to their number rather than their violation of social distancing behaviors. On the other hand, although risk perception and self-restriction recommendations can induce social distancing behaviors, their impact and effectiveness appear to be largely weakened during Japan's second state of emergency.

Conclusions: This study provides a timely reference for policymakers about the current situation on how different demographic groups adopt social distancing behaviors over time. On the one hand, the age-dependent disparity requires more nuanced and targeted mitigation strategies to increase the intention of elderly individuals to adopt mobility restriction behaviors. On the other hand, considering that the effectiveness of policy implementations requesting social distancing behaviors appears to decline over time, in extreme cases, the government should consider imposing stricter social distancing interventions, as they are necessary to promote social distancing behaviors and mitigate the transmission of COVID-19.

XSL·FO
RenderX

Introduction

Background

The rapid global prevalence of COVID-19 has caused an unprecedented public health crisis. Currently, social distancing measures require avoiding unnecessary physical contact and remain the primary public health strategy for mitigating the spread of COVID-19. Several studies have indicated that the transmission rate and death rate seem correlated with how firmly social distancing was implemented [1,2]. However, due to the substantial economic and psychological cost [3,4], social distancing measures are not necessarily accomplished by the whole population. Indeed, previous studies have suggested that protective behaviors are associated with demographic factors, including gender and age [5]. In this sense, variation of infectious cases among demographic groups might be attributed to the varying levels of social distancing behaviors across demographic groups [6-9]. Investigating and understanding the variation in social distancing behaviors across demographic groups can improve the design, implementation, effectiveness, and equity of public health policies, which can reduce the spread of infection and limit the outbreak.

Compliance with social distancing measures during the COVID-19 pandemic has received increasing attention. A primary line of research has conducted surveys to assess individuals' social distancing behaviors. However, the inherent limitations of surveys determine that they could only capture the condition in a manner limited to relatively coarse areal units or a short-term period. Thus, these studies can only provide insights into 1 point or a short-term situation of social distancing behaviors, and the real-time change in behaviors at higher spatial and temporal resolutions cannot be fully quantified. As the spread of COVID-19 has been due to a long-lasting pandemic, social distancing behaviors should be considered a dynamic process that evolves and shifts in individuals' perceptions and in policies. Thus, an established mechanism for social distancing behaviors must be examined from a long-term perspective. In this sense, how social distancing behaviors shift in response to different periods remains an open question.

The availability of large-scale and fine-grained aggregated mobility data has allowed researchers to observe the dynamics of social distancing behaviors at high spatial and temporal resolutions [10-12], which can naturally serve as an appropriate assessment to produce a more scalable, long-term analysis. Studies have used mobility data to observe the patterns of human activities during the COVID-19 pandemic and assess the effectiveness of social distancing measures [13-15]. However, most studies have primarily used mobility data to derive coarse information about the estimation of population flow for mathematical simulation and how demographic factors related to mobility have not been fully investigated, as the reliable demographic-specific mobility data remain scarce.

Against these backgrounds, this study explored the social distancing behaviors among various demographic groups over time by using fine-grained mobility tracking data combined with demographic information. More specially, to assess social distancing behaviors by evaluating human mobility data, we focused on the mobility population in entertainment venues. During the current pandemic, officials have strongly encouraged individuals to reduce the frequency of their visits to nonessential leisure establishments, such as restaurants and bars. Notably, an increasing number of infections linked to these settings have been observed, indicating that visiting entertainment venues increases individuals' risk of infection. Therefore, we considered that visiting entertainment venues is a typical violation of social distancing measures and would be an appropriate proxy for social distancing behaviors.

In addition, this study specified the voluntary response and policy-induced response to provide a more comprehensive understanding of social distancing behaviors.

First, protection motivation theory suggests that individuals primarily tend to adopt voluntary protective behaviors, including maintaining social distancing, because of their desire to avoid the risk and adverse outcomes of infection [16-18]. Risk perceptions have been important drivers of individuals' social distancing behaviors during the COVID-19 pandemic. As the perceived susceptibility and perceived severity of the disease can vary across demographic groups [5] and shift over time, it is reasonable to assume there are also demographic differences and time variations in the compliance with social distancing measures.

In addition, public health policy implementation can significantly affect the extent of compliance with social distancing measures. Doubtlessly, beyond the voluntary compliance, policies implemented by governments could accelerate and strengthen compliance with social distancing measures. However, a more nuanced investigation is essential to better understand the impact of public health policy implementation. On the one hand, the effectiveness of the policies is largely dependent on public acceptance and obedience, which might be varied across demographic groups. Therefore, the impact of public health policy on social distancing behaviors can also vary across demographic groups. Governments need to ensure that the policies target all demographic groups, while considering that different demographic groups might not equivalently respond to the policies. On the other hand, government-initiated interventions have been relatively short lived as their implementation can cause substantial social-economic costs. Notably, the resurgence of the epidemic after the lifting of strict social distancing measures may again pose a severe threat and force policymakers to impose stringent social distancing measures repeatedly [19,20]. In this context, the implementation of policy interventions may be enforced multiple times, while the actual effect of these policies on social distancing behaviors during the different phases has not been fully investigated.

In summary, this study aims to address 2 research questions (RQs):

- RQ1: Can an increase in the cases of infection lead to a decrease in the frequency of visits to entertainment venues? Does their relationship vary across demographic groups and different periods of the COVID-19 pandemic?
- RQ2: How does policy intervention affect visits to entertainment venues? Does its impact vary across

XSL•FO
RenderX

demographic groups and different periods of the COVID-19 pandemic?

Research Case

Our focus is on the Tokyo metropolitan area, 1 of the areas most affected by COVID-19. As of March 31, 2021, Tokyo already had 120,986 confirmed COVID-19 cases, which resulted in 1804 deaths. Figure 1 presents the changes in the COVID-19 daily new confirmed cases in Tokyo.

The first case of COVID-19 in Japan was confirmed on January 24, 2020, while confirmed infectious cases remained relatively limited in the following few weeks. However, after the number of cumulative confirmed cases exceeded 100 on February 21, 2020, the spread of the virus began to progress rapidly. On March 26, 2020, the Tokyo government officially released a

policy calling for self-restriction, including closure, shortened business hours, and limited capacity in entertainment venues in the Tokyo metropolitan area. Moreover, the government declared the state of emergency to further induce self-restriction behaviors. Although the rate of new infections temporarily decreased during the implementation of the policies and the national emergency response ended on May 25, 2020, the transmission of COVID-19 continued. The number of daily new confirmed cases has been increasing since July 2020 and eventually peaked in December 2020. To mitigate and interrupt the spread of COVID-19, the Japanese government implemented the second state of emergency from January 8 to March 21, 2021, to limit individual mobility and promote social distancing. As indicated by the labels in Figure 1, the targeted period can be separated into several time windows according to these key time points.

Figure 1. Timeline of COVID-19 prevalence in the Tokyo metropolitan area.
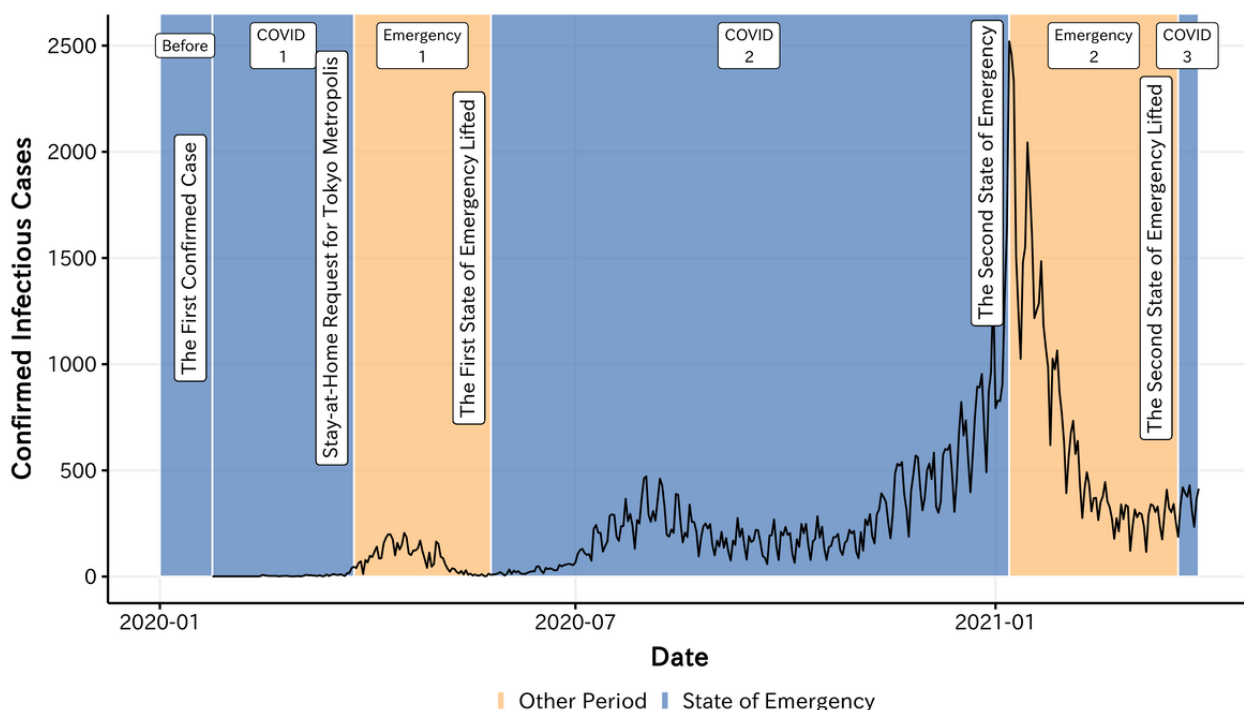
To balance the benefits and costs of social distancing measures, the Japanese government imposed them mainly as recommendations; thus, citizens were expected to voluntarily comply by, for example, refraining from performing outdoor activities and avoiding gatherings. In this sense, because the success of social distancing behaviors depends on spontaneous public acceptance and compliance, it is imperative to investigate not only which demographic groups are more or less likely to accept and comply with the restrictions but also how the compliance changes across time. It should be noted that although social distancing measures were still accompanied by voluntary compliance without legal penalties even during the period of the state of emergency, more restrictive policies, including canceling events, closing down nonessential business facilities, and reducing business hours, were implemented to minimize physical contact. Nevertheless, as a particular case, the policy implementation's effect that primarily relies on voluntary compliance remains unclear. Particularly, because Japan

experienced multiple waves of the rapid spread of COVID-19 infection and states of emergency, the impact of policy implementations might vary over time. Investigating how compliance with social distancing measures was affected by policy implementation throughout the period of the COVID-19 pandemic can provide essential insights into clarifying whether a "mild lockdown" [21] policy is the appropriate intervention strategy for guiding social distancing behaviors.

Methods

Study Design

This study combined several data sources, including mobile tracking data and geographical statistics, to estimate mobility dynamics in entertainment venues across demographic groups. First, we established a definition of entertainment venues based on geographical statistic data related to the constitution of the facility and the workers' type in the specific area. After

specifying entertainment venues, we used the visiting population of these entertainment venues to construct a mobility index as the proxy of social distancing behaviors. Finally, we combined the mobility data with demographic information and investigated how social distancing behaviors varied across demographic groups and shifted over time.

The key data sources and definitions are introduced next.

Mobile Tracking Data

In this study, we used mobility data provided by DOCOMO Insight Marketing, Inc [22], 1 of the biggest telecom companies in Japan. Using the cell towers, the locations of the individuals were recorded on an hourly basis and aggregated as the estimated population with a 500 m grid cell. Particularly, based on the profile of the mobile subscribers, geolocation information was aggregated into demographic groups. Furthermore, the original data were preprocessed through extrapolating estimation. Thus, mobility population can be used to reflect the actual condition of mobility without bias in the adoption rates of NTT DOCOMO mobile terminals by age group, gender, and residential area [22,23]. Mobile tracking data provide high-spatial, longitudinal, and demographic-specific mobility populations for examining how mobility patterns vary across demographic groups over time. We focused on the mobility data recorded in the Tokyo metropolitan area from January 1 to March 31, 2021. This period covers the time before and during the COVID-19 outbreak in Japan, allowing for a comprehensive analysis of mobility behaviors over time.

Specification of Entertainment Venues

This study focused on the mobility in entertainment venues in which most of the space was used for leisure, for example, restaurants and bars. To specify major entertainment venues in the Tokyo metropolitan area, we used granular land-use data from the Statistics Bureau of Japan (2016) that includes area-feature information on the composition of establishments and workers in 500 m grid cells by industry. In our case, we defined the industrial divisions "accommodation, eating, and drinking services" and "living-related, personal, and amusement services" as an entertainment-related category and assumed that areas where the proportion of establishments and workers was high could be considered "entertainment venues." More specifically, the eligibility criteria of a 500 m grid cell that identified as an entertainment venue were (1) the number of entertainment category–related establishments was >100, (2) the number of entertainment category–related workers was >500, (3) the proportion of entertainment category–related establishments was >0.4, and (4) the proportion of entertainment category–related workers was >0.4.

Mobility Population and Mobility Index

As outcome measures, this study assessed social distancing behaviors based on metrics related to the mobility population. The mobility population indicates the volume of the population tracked in entertainment venues, which could be easily computed by matching the specified 500 m grid entertainment venues and the corresponding mobility data. However, the mobility population cannot be directly applied to estimate the dynamic of social distancing behaviors, as the dynamic of the mobility population may have an inherent seasonal pattern. Instead, a mobility index was computed to reflect the extent to which mobility changed because of the COVID-19 pandemic during the study period, which we compared with the baseline. More specifically, we used the mobility population in entertainment venues in 2019 as a baseline that represents the level of the mobility population without the effect of the pandemic. Next, we compared the daily time series of the aggregated mobility population since 2020 with this baseline to compute the mobility index, which controlled potential seasonality factors that may affect mobility patterns other than those of the pandemic. Formally, for a specific demographic group "i," distinguished by age and gender, we used the following formula to estimate the mobility index since 2020:

where r = {r_1, r_2, . . . r_n} denotes a list of entertainment venues and "t" denotes the date of mobility population tracked. Here, is the mobility population in all defined entertainment venues of demographic group "i" during a specific date, and is the mobility population of the corresponding date in 2019 as a baseline.

Beyond the macro seasonal variation, such estimated mobility index might still be biased due to the inherent variation in the mobility population during the week.

As shown in Figures 2A and 2B, before the pandemic, the mobility population in

entertainment venues was high during weekends in 2019, while since 2020, the magnitude of variation decreased due to the pandemic. Since the comparison with the baseline was according to the date, for example, when the specific date is a weekday in 2020 but a weekend in 2019, the mobility index would be estimated biasedly smaller and vice versa. To exclude the day-of-the-week effects, we further processed the mobility index by computing the 7-day moving average as follows:

As shown in Figure 2C, the dynamic of the original mobility index seems to exhibit instability due to the day-of-the-week effects. Through the computation of the moving average, the dynamic of the estimated mobility index tends to be smooth and can still represent the overall trend of the original mobility index. As statistical analysis can be sensitive to the bias in the original mobility index, we used the moving average of the mobility index to compute social distancing behaviors in the following analysis.

The main objective of this study was to provide a comprehensive understanding of how voluntary and policy-induced social distancing behaviors shift over time across demographic groups. We conducted the analysis using 2 methods.

Figure 2. (A) Weekly variation in the mobility population during 2019. (B) Weekly variation in the mobility population since 2020. (C) Comparison of the original mobility index and the 7-day moving average of the mobility index.
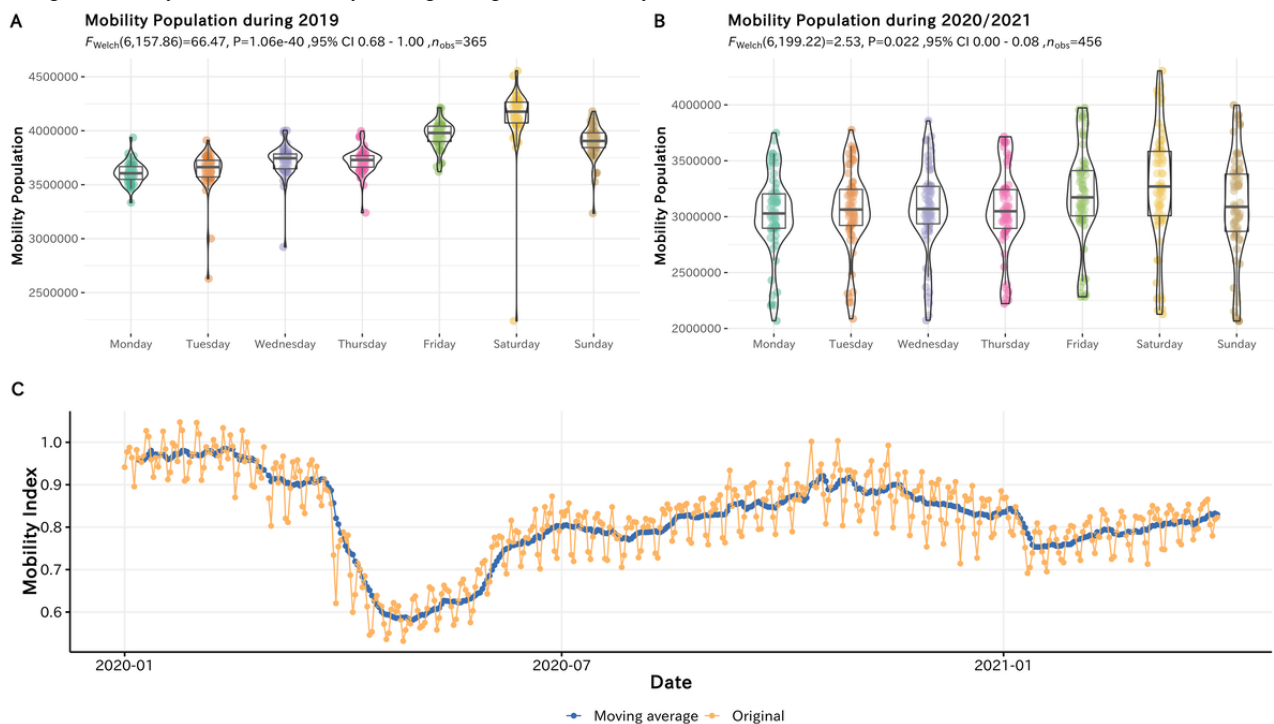
Voluntary Social Distancing Behaviors

To investigate voluntary social distancing behaviors, we focused on the temporal relationship between the prevalence of infectious and the estimated mobility index of entertainment venues. We assumed that individuals' voluntary compliance with social distancing behaviors was based on their risk perception, which is related to the recent prevalence of infectious. In practice, we used a cross-correlation function (CCF) to examine the association between time-lagged daily cases of infection in Tokyo and the moving average of the mobility index. A CCF measures chronological relations between the time series "x" and the time-shifted time series "y." In our case, the estimated coefficient of the CCF can be interpreted as a metric that describes how recent cases of infection affected mobility in entertainment venues. In practice, we incrementally shifted the daily cases of infection back forward from 0 to 7 days and computed the CCF between the mobility index and the daily cases of infection at different lags.

Policy-induced Social Distancing Behaviors

Beyond the social distancing behaviors induced by the risk perception, we also focused on how policy intervention induced the change in social distancing behaviors. More specifically, we focused on the 2 post-policy-intervention periods in the Tokyo metropolitan area—from March 26 to May 5, 2020, and from January 1 to March 22, 2021—and then used the Bayesian structural time series (BSTS) model [24] to dynamically investigate the policy-induced social distancing behaviors. BSTS models can simulate counterfactual trends based on the model training on the pretreatment time series, that is, predicted counterfactual series that would have occurred in a virtual counterfactual scenario with no intervention. Subsequently, the causal effect of the intervention can be determined by computing the pointwise relative impact, and the cumulative causal impact can be assessed by comparing the real postintervention observed series and the predicted postintervention observed series.

Our research design used the BSTS models to investigate how policy intervention affects the estimated mobility index in entertainment venues. In addition, BSTS models are allowed to incorporate covariates likely to affect the outcome of interest to control for spurious effects and unobservable dynamics. Notably, time-varying covariates are assumed to be unaffected by the effects of intervention treatment. In practice, we assumed that mobility in the entertainment venues was likely to be influenced by weather conditions. Thus, we incorporated the mobility index, temperature, precipitation, and wind velocity into BSTS models that integrate weather conditions to fit the trend of the mobility index.

Analysis Framework

In summary, our analyses are organized as follows.

First, we extracted the mobility population within the entertainment venues of each demographic group to access social distancing behaviors.

Subsequently, we computed the mobility index to capture the dynamics of social distancing behaviors during the COVID-19 pandemic. On the one hand, we used the mobility population for 2019 as a baseline to compute the mobility index and to control for seasonality variation and the imbalanced population of the demographic groups. On the other hand, we computed the 7-day moving average of the mobility index to investigate the dynamics of social distancing behaviors and to control for the day-of-the-week effect.

Next, with the estimated mobility index, we used the CCF to examine the effect of the increase in infectious cases on social

distancing behaviors and the degrees of this influence across the demographic groups and periods of the COVID-19 pandemic.

Moreover, we used the BSTS model to investigate the effect of the state of emergency in Japan on social distancing behaviors. Particularly, the model was computed for 2 states of emergency. In this manner, the study provided insight into changes in the policy-induced social distancing behaviors of different demographic groups during the COVID-19 pandemic in Japan.

Results

Validation of the Estimated Mobility Index in Entertainment Venues

As a quality control of our estimation, we cross-validated the estimated mobility index in entertainment venues by comparing it with the time series data provided by Google Mobility Reports [25], which suggest the extent to which mobility in a certain category of places changes compared with the baseline. As shown in Figure 3, we compared the mobility in the "retail and recreation" category in Tokyo and our estimated mobility index in the entertainment venues for each day. Both the original mobility index (R^2=0.77, 95% CI 0.73-0.80, P<.001) and the moving average of the mobility index (R^2=0.87, 95% CI 0.85-0.89, P<.001) were highly correlated with the outcome of Google Mobility Reports, indicating that our estimation of the mobility that occurred in entertainment venues was generally reasonable.

Figure 3. Correlation between the estimated mobility index and Google's mobility report on retail and recreation.
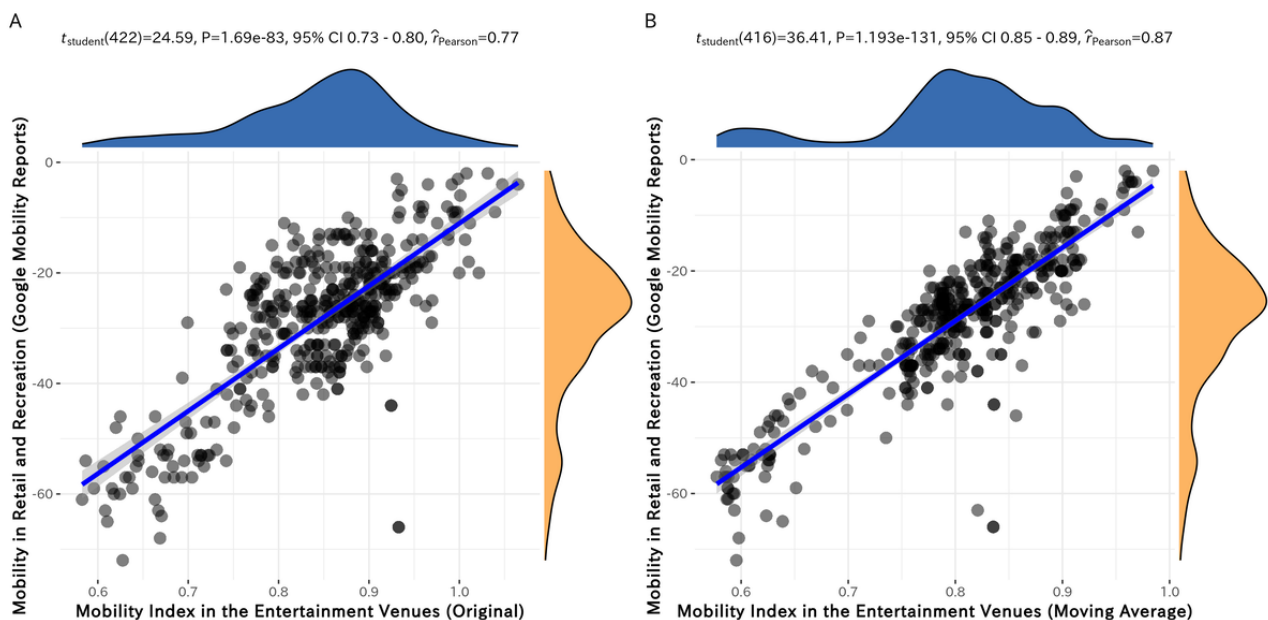
General Mobility Dynamics by Gender and Age

This section presents the general mobility dynamics of each demographic group. Vertical dashed lines indicate the period in which the state of emergency was implemented in Tokyo metropolitan areas.

Figure 4 presents the dynamics of the mobility population in the entertainment venues by gender and age. On the one hand, by comparing the mobility pattern from the perspective of gender, we found that the general mobility population of males is higher than that of females. Particularly, among individuals aged 30-70 years, the mobility population of males was significantly higher than that of females [8]. On the other hand, from the perspective of age, the mobility populations of individuals in their twenties were significantly higher than those of the elderly, which implies that individuals in their twenties constitute the majority of the population in entertainment venues. For details on the comparison of mobility populations among demographic groups, please refer to Multimedia Appendix 1.

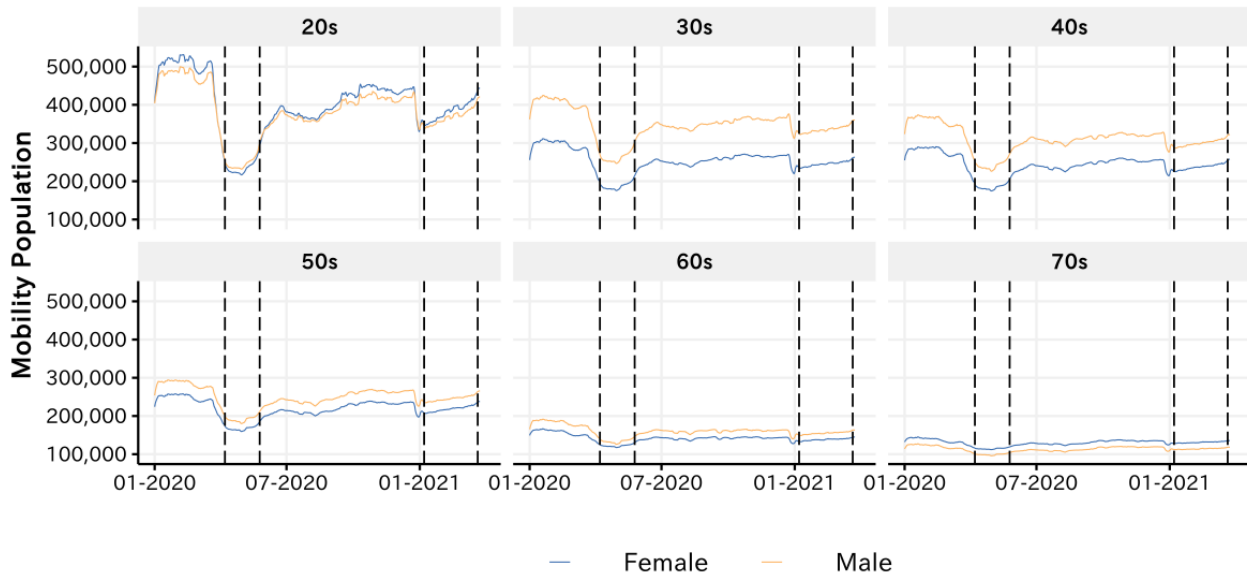**Figure 4.** Dynamics of the mean average mobility population by age and gender.

Figure 5 presents the dynamics of the estimated mobility index in entertainment venues by gender and age. Generally, we observed a similar dynamic pattern across demographic groups: the mobility index significantly decreased across demographic groups after the outbreak of the COVID-19 pandemic in Japan. For a detailed comparison of the mobility index between the pre-COVID-19 period and the COVID-19 pandemic, please refer to Multimedia Appendix 2. Also consistent was that when the policy interventions were implemented, a significant decrease in the mobility index was observed, followed by a

rebound to the pre-policy-intervention after lifting of the policy interventions. In general, no significant differences were observed in the estimated mobility index between males and females, whereas the dynamics of the estimated mobility index appeared to vary across age groups. Particularly, after the outbreak of the COVID-19 pandemic, the mobility index of individuals in their twenties appeared to decrease more significantly than that of the elderly, which implies that a large proportion of youths were adopting social distancing behaviors.

Figure 5. Dynamics of the mean average mobility index by age and gender.
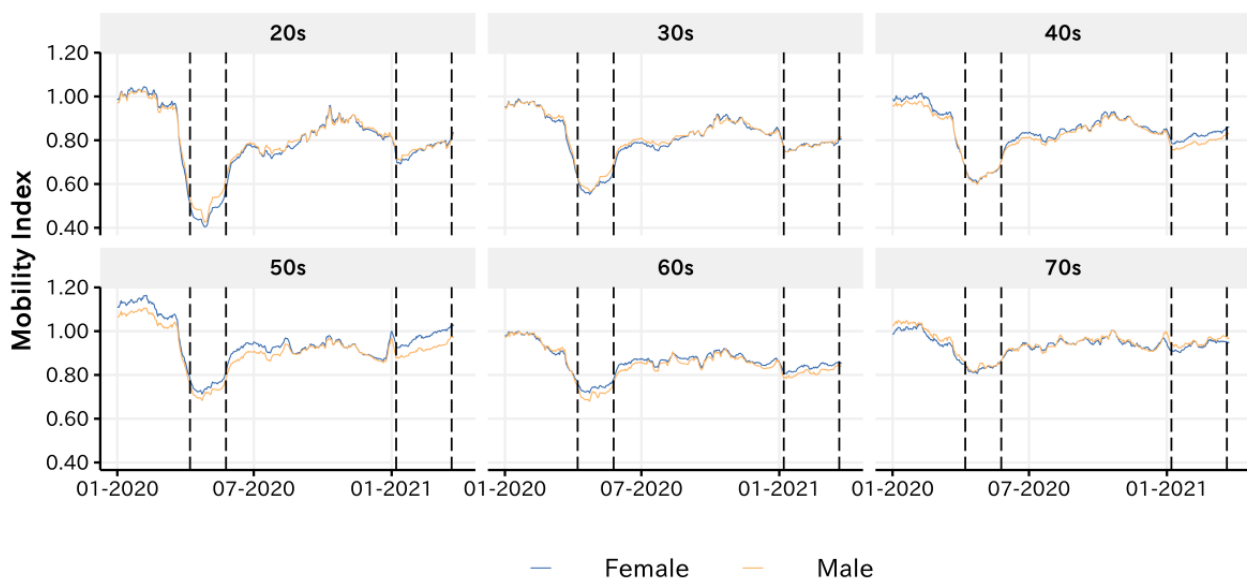
Voluntary Social Distancing Behaviors

A typical driver of voluntary social distancing behaviors was the risk perception induced by the increasing number of infections. Here, we used the CCF to examine the association between the prevalence of infections and the dynamics of voluntary social distancing behaviors across demographic groups over time. Specifically, negative correlation coefficients indicated that the increase in the cases of infection could have

led to the decrease in the mobility index in entertainment venues, which we interpreted as the magnitude of voluntary social distancing behaviors.

According to the period of the COVID-19 pandemic in Japan defined earlier, we computed the CCF for each demographic group during the different periods of the COVID-19 pandemic.

Although no systematic differences in voluntary social distancing behaviors were observed among demographic groups (please refer to Multimedia Appendix 3 for detailed analysis results), we found that the pattern of voluntary social distancing behaviors shifted during different periods of COVID-19 pandemic. Figure 6 presents the cross-correlation between the number of cases of infection and the mobility index during the different periods of the COVID-19 pandemic in Japan. In Figure 6, the horizontal dashed lines indicate the boundary of white noise, and the values of the CCF must lie beyond the interval to be significant. We observed that during the COVID-1 period, the number of infectious cases was negatively correlated with the mobility index at lags from 0 to 2 days, which indicated that

the recent increasing prevalence of infections can lead to a decline in visiting entertainment venues. However, statistical significance of association was not observed during both the Emergency-1 period and the COVID-2 period. This finding indicates that the increasing number of cases of infection rarely affects visits to entertainment venues, that is, the magnitude of voluntary social distancing behaviors has gradually decreased during these periods. During the Emergency-2 period, the number of cases of infection and mobility index became strongly correlated again, indicating that the second declaration of the state of emergency significantly activated the voluntary social distancing behaviors.

Figure 6. Cross-correlation between the number of cases of infection and the mobility index during the periods of the COVID-19 pandemic. CCF: cross-correlation function.
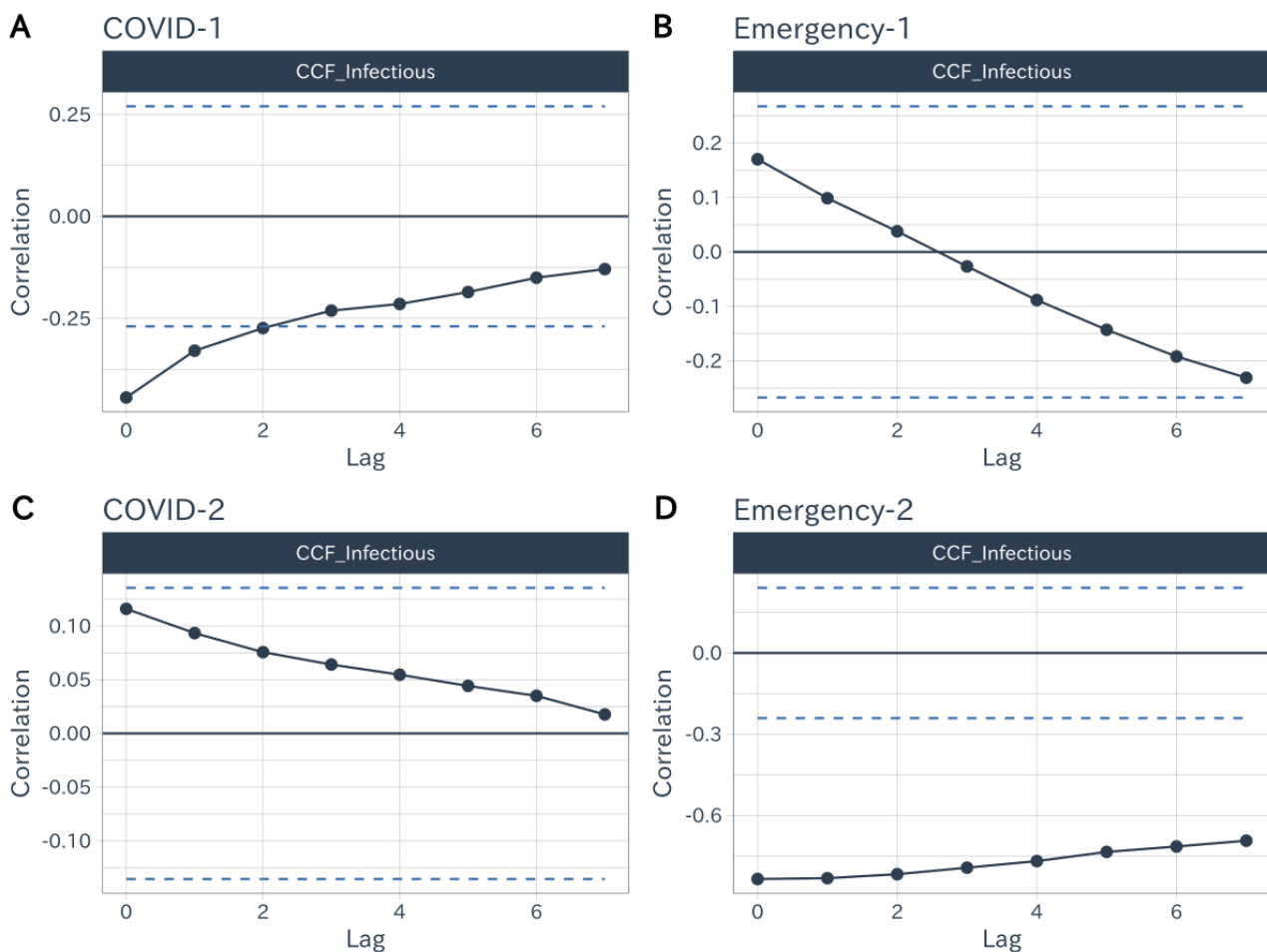
Policy-induced Social Distancing Behaviors

To examine the impact of policy intervention, we used the BSTS method to predict the counterfactual mobility dynamics in entertainment venues and could thus quantitatively analyze how policy interventions affect social distancing behaviors. More specially, we set 2 types of models to evaluate the impact of 2 policy interventions, respectively: (1) the COVID-1 period as the pretreatment period and the Emergency-1 period as the posttreatment period and (2) the COVID-2 period as the

pretreatment period and the Emergency-2 period as the posttreatment period.

Figure 7 demonstrates how the BSTS model made the predictions and constructed the counterfactual predictions for accessing the interventions' impact. More specifically, the vertical dashed line denotes the day that the state of emergency was released, the solid series denotes the real dynamics of the mobility index, and the dashed series indicates the counterfactual trends that would have been observed without intervention.

Figure 7. Time series of the Bayesian structural time series (BSTS) model.
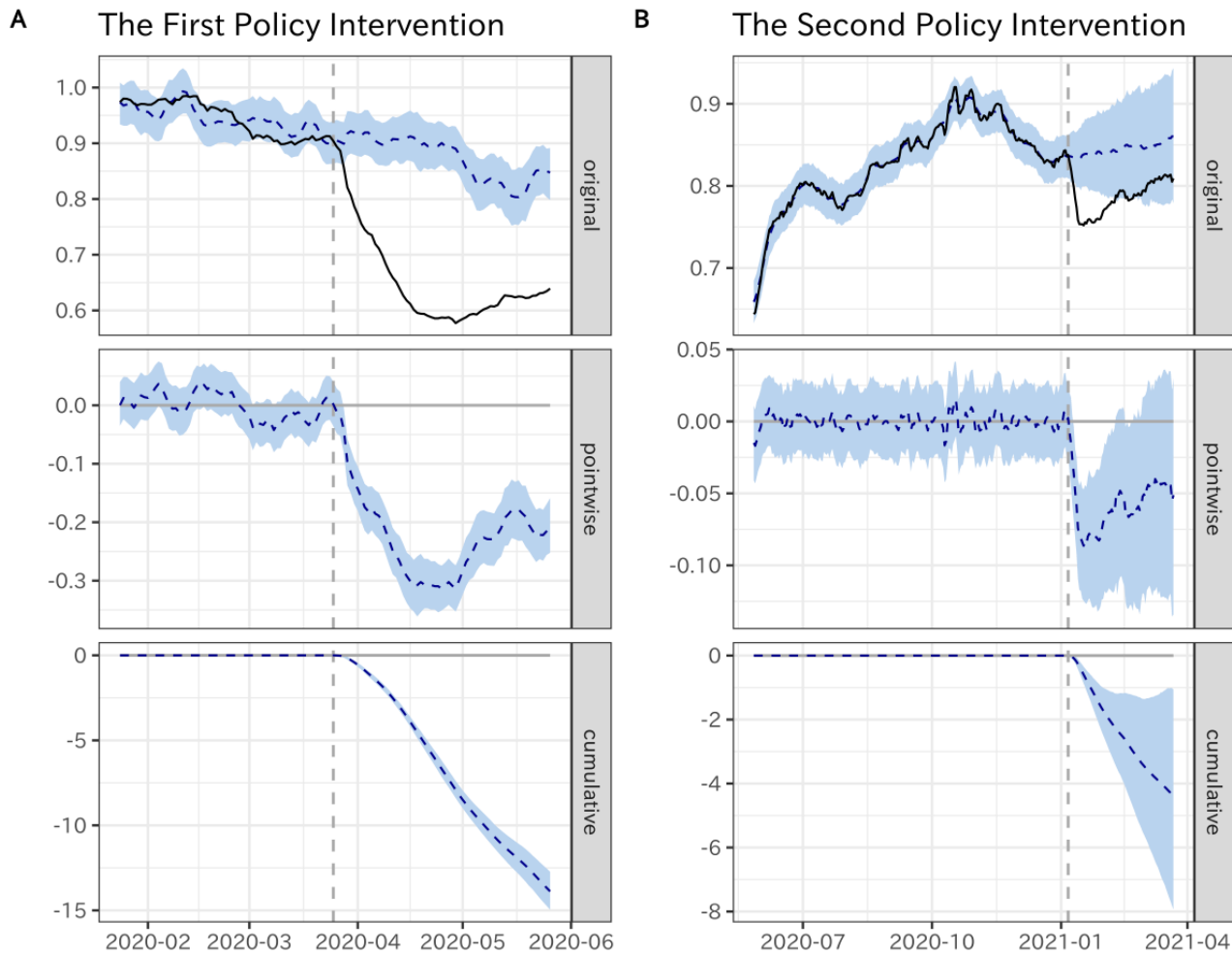
For both computations, we found that during the pretreatment period, the BSTS model successfully captured the long-term trend and the fluctuations caused by the covariates. After the intervention, we observed considerable differences between the 2 series, which were summarized in terms of pointwise differences and cumulative differences.

In practice, we measured the impact of the 2 policy implementations on different demographic groups separately. Figure 8A presents the relative impact among different demographic groups. In Figure 8A, the vertical dashed line indicates the general impact of the 2 policy interventions that were computed on the whole population, which can serve as the baseline of policy impact.

Generally, as the estimated coefficients were lower than 0 across demographic groups, the impact of policy implementations was applicable to the whole population. However, we found that the magnitude of the effect appeared to vary by demographic groups over time.

On the one hand, for the first policy intervention, we observed a considerable decrease in the mobility index among most demographic groups. Especially, the mobility index for females aged 20-29 years decreased by 42.73% and that of males aged 20-29 years decreased by 40.19%, which is significantly higher than elderly individuals. These findings indicate that the policy interventions had the most significant impact on youths.

On the other hand, for the second policy intervention, although it also caused the visiting population to entertainment venues to slightly decrease, the magnitude of the impact substantially dropped compared to the former policy intervention among all demographic groups. That is, although the state of emergency was announced again to restrict nonessential outdoor activity, many people maintained their regular behavior pattern rather than complying with the social distancing requests, and the impact of such policy implementation largely declined.

Beyond the general magnitude of the impact, because the length of the 2 policy intervention periods was similar, we compared and investigated how the dynamic pattern of impacts varied between the 2 policy interventions.

Figure 8. (A) Relative impact of 2 policy interventions across demographic groups. (B) Comparing the pointwise impact between the 2 policy interventions. (C) Comparing the cumulative impact between the 2 policy interventions.
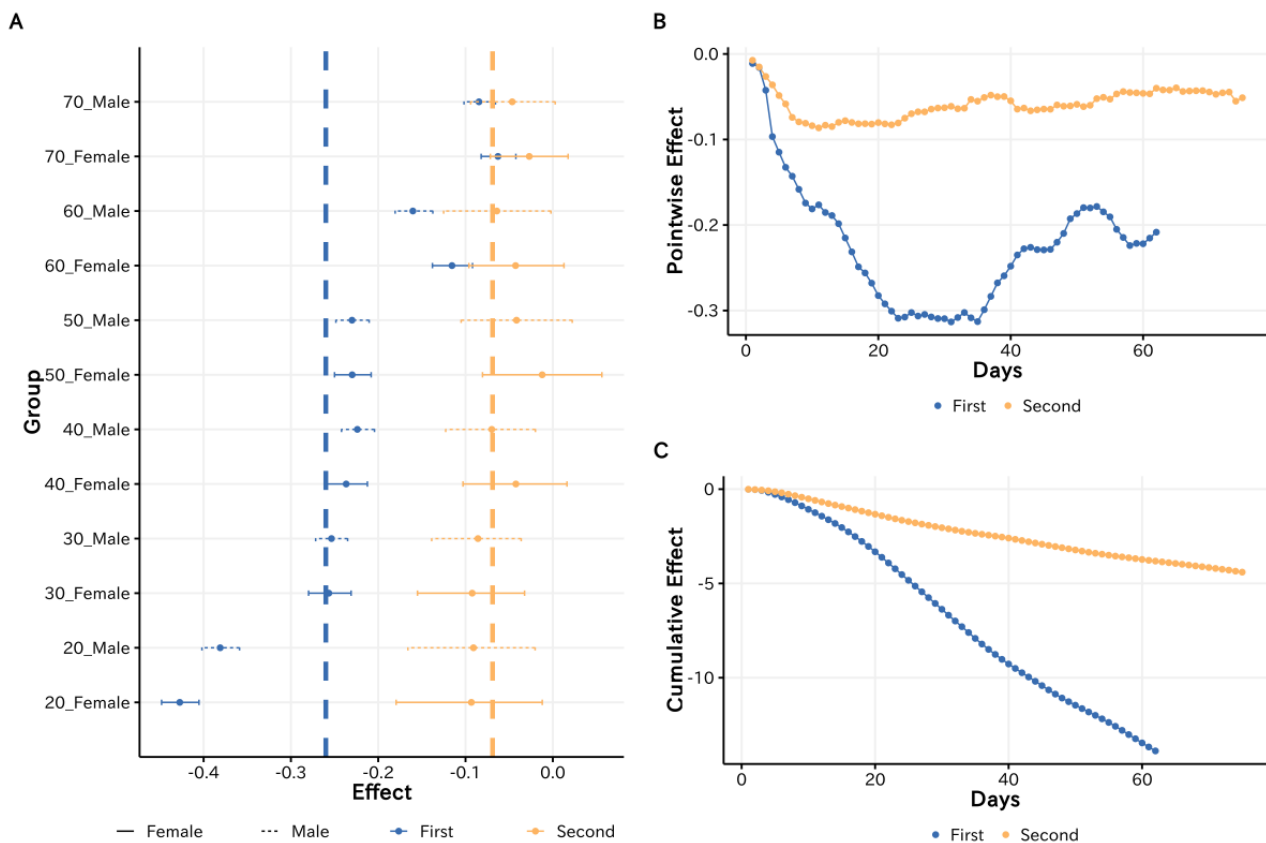
We present Figures 8B and 8C to compare the dynamics of the pointwise impact and the cumulative impact, respectively, between the 2 policy interventions. For both 2 policy interventions, we observed that the implementation of the policy interventions consistently led to a decrease in visits to entertainment venues during the initial few days, while the impact of policy interventions decreased in the following days. In other words, the policy-induced compliance with the social distancing measures tended to have a transient phase—with a significant reduction in mobility for the initial policy implantation—and the impact of the policy on compliance willingness substantially decreased over time. Here, we specifically compared transient trends between the 2 policy interventions. On the one hand, as shown in Figure 7B, the time point that the impact turned to decrease was much earlier for the second policy intervention than for the first policy intervention. On the other hand, as shown in Figure 7C, the gradient of the cumulative impact was much higher for the first policy intervention than for the second policy intervention. In summary, these comparisons suggest that the impact of the first policy intervention on social distancing behaviors was not only significantly greater but also tended to be much longer-lasting compared to that of the second policy intervention.

Discussion

Principal Findings

In this study, we used mobility data to investigate how social distancing behaviors vary among demographic groups. The disparity and dynamics of social distancing behaviors driven from our analysis can have critical implications for optimal disease control policy design and implementation.

First, our findings demonstrate distinct patterns of social distancing behaviors and their dynamics across age groups. More specifically, we found that the population in entertainment venues comprised mainly individuals aged 20-40 years, which implies that this age group could be exposed to a higher risk of infection in entertainment venues. However, the larger amount of population did not necessarily suggest that the individuals in this age group are more likely to violate the restrictions. On the one hand, based on the dynamics of the estimated mobility index, among the age groups, the extent of reduction in the frequency of visiting entertainment venues during the pandemic was generally higher among younger individuals, particularly individuals aged 20-30 years. On the other hand, by investigating the impact of policy interventions, the rates of acceptance and compliance with the social distancing policy interventions were also higher among younger individuals. From this perspective, the increasing contribution of the youth to the spread of infection should be more likely attributed to their size instead of their refusal to observe social distancing behaviors. Indeed, the adoption of social distancing behaviors is less likely induced by policy implementations among the elderly. As the COVID-19 pandemic progresses, increased nuanced and targeted responses are required to effectively control the prevalence of the infection. Given the age-dependent disparity, governments should tailor mobility restrictions to the targeted population. Especially, the existing policy implementations are seemingly inefficient for the elderly, who are more susceptible to the severe symptoms

of and mortality posed by COVID-19. Thus, target mitigation strategies might be necessary to increase the intention of elderly individuals to adopt mobility restriction behaviors.

Second, our analysis also provides insights into the dynamics of voluntary social distancing behaviors over time. Our proposed mechanism assumes the connection between voluntary social distancing behaviors and the increasing number of infections mediated by risk perception. Accordingly, we propose that the dynamics of voluntary social distancing behaviors can be explained by the dynamic of risk perceptions. Our results indicate that in the initial phase of the COVID-19 pandemic, the increasing number of cases of infection could have led to the decrease in entertainment venue visiting; thus, voluntary social distancing behaviors have resulted in important outcomes in terms of reducing unnecessary physical contact during this period. However, with the progress of the pandemic, the significance of such association declined, which indicates that the risk perceptions about COVID-19 have decreased over time. This scenario may be attributed to the decreased adherence of the public to social distancing measures and vigilance toward COVID-19 [26]. Moreover, we find that policy interventions can strengthen risk perceptions. Specifically, although voluntary social distancing behaviors largely diminished during the COVID-2 period or the second state of emergency, policy intervention appeared to increase the awareness of the severity of the pandemic and concerns regarding COVID-19, leading to an increase in voluntary social distancing behaviors. In this sense, policymakers should continue to alert the public about the risk of COVID-19 in order to promote voluntary social distancing behaviors.

Third, our results indicate the importance of implementing the public health policy promptly to limit the spread of the COVID-19 infection. Quantifying the impact of policy interventions is crucial for policymaking. Here, 2 insights deserve emphasis. On the one hand, although the social distancing interventions in Japan were less strict than those in some other countries, they still significantly promoted social distancing behaviors under the implementation of a state of emergency, which is in with previous investigations [27], although the adoption of social distancing behaviors resurged and then gradually resumed to the normal level after lifting the policy interventions. On the other hand, our results warn policymakers that the effectiveness and impact of self-restriction recommendations appeared to decrease in response to the second wave of COVID-19. Particularly, in the second state of emergency in Japan, the magnitude of the reduction in the visiting flow to entertainment venues was limited compared to the first state of emergency. Furthermore, the initial, strong impacts could only last for a short time and could quickly enter the decreasing phase.

Limitations

The findings of this study should be carefully considered in the context of its 2 main limitations.

The first limitation is that we focused on the mobility flow in entertainment venues as a proxy to estimate social distancing behaviors; however, visiting entertainment venues is only 1 aspect related to compliance with social distancing measures. Nevertheless, this study presented an example of how to integrate aggregated mobility data with geological statistics data. Hence, we suggest that in further research, the proposed analysis framework that integrated mobility data and geographical statistics data be applied to monitor other aspects of social distancing behaviors. This research direction would be promising in attempts to extend the methodology to specify other types of locations (eg, residential areas or business districts); subsequently, mobility data could be used to investigate social distancing behaviors from the perspective of compliance with stay-at-home orders or remote work orders. Another direction for further research would be to provide a more comprehensive set of insights into social distancing behaviors.

The second limitation is that the implication related to policy-induced social distancing behaviors is based on the scenario of the policy implementation in Japan, where social distancing measures were accomplished through spontaneous cooperation. Thus, caution should be exercised when interpreting the implications, because it may not be applicable to other regions or countries, where the enforcement of social distancing was stricter than that in Japan.

Conclusion

Given the costs of the enforced policy, many countries have decreased the stringency of the containment policy. Thus, voluntary social distancing behaviors are expected to play a critical role in future responses to the COVID-19 pandemic, even for countries that mainly relied on the enforced containment policy at the initial phase. From this perspective, implications derived from Japan could be generalized to other countries and serve as guidance for the effective induction of voluntary social distancing behaviors to combat the long-term COVID-19 pandemic.

Conflicts of Interest

None declared.

XSL•FO

RenderX

Multimedia Appendix 1
Comparison of mobility population across demographic groups.
[DOCX File , 5589 KB - medinform_v10i3e31557_app1.docx]

Multimedia Appendix 2
Comparison of mobility index before and during the COVID-19 pandemic.
[DOCX File , 465 KB - medinform_v10i3e31557_app2.docx]

Multimedia Appendix 3
Cross-correlation across demographic groups and time periods.
[DOCX File , 330 KB - medinform_v10i3e31557_app3.docx]

References

1. Islam N, Sharp SJ, Chowell G, Shabnam S, Kawachi I, Lacey B, et al. Physical distancing interventions and incidence of coronavirus disease 2019: natural experiment in 149 countries. BMJ 2020 Jul 15;370:m2743 [FREE Full text] [doi: 10.1136/bmj.m2743] [Medline: 32669358]

2. Courtemanche C, Garuccio J, Le A, Pinkston J, Yelowitz A. Strong social distancing measures in the United States reduced the COVID-19 growth rate. Health Aff (Millwood) 2020 Jul 01;39(7):1237-1246. [doi: 10.1377/hlthaff.2020.00608] [Medline: 32407171]

3. Brooks SK, Webster RK, Smith LE, Woodland L, Wessely S, Greenberg N, et al. The psychological impact of quarantine and how to reduce it: rapid review of the evidence. Lancet 2020 Mar;395(10227):912-920. [doi: 10.1016/s0140-6736(20)30460-8]

4. Holmes EA, O'Connor RC, Perry VH, Tracey I, Wessely S, Arseneault L, et al. Multidisciplinary research priorities for the COVID-19 pandemic: a call for action for mental health science. Lancet Psychiatry 2020 Jun 15;7(6):547-560 [FREE Full text] [doi: 10.1016/S2215-0366(20)30168-1] [Medline: 32304649]

5. Bish A, Michie S. Demographic and attitudinal determinants of protective behaviours during a pandemic: a review. Br J Health Psychol 2010;15(4):824. [doi: 10.1348/135910710x485826]

6. Weill JA, Stigler M, Deschenes O, Springborn MR. Social distancing responses to COVID-19 emergency declarations strongly differentiated by income. Proc Natl Acad Sci U S A 2020 Aug 18;117(33):19658-19660 [FREE Full text] [doi: 10.1073/pnas.2009412117] [Medline: 32727905]

7. Al-Hasan A, Yim D, Khuntia J. Citizens' adherence to COVID-19 mitigation recommendations by the government: a 3-country comparative evaluation using web-based cross-sectional survey data. J Med Internet Res 2020 Aug 11;22(8):e20634 [FREE Full text] [doi: 10.2196/20634] [Medline: 32716896]

8. Clark C, Davila A, Regis M, Kraus S. Predictors of COVID-19 voluntary compliance behaviors: an international investigation. Glob Transit 2020;2:76-82 [FREE Full text] [doi: 10.1016/j.glt.2020.06.003] [Medline: 32835202]

9. DeFranza D, Lindow M, Harrison K, Mishra A, Mishra H. Religion and reactance to COVID-19 mitigation guidelines. Am Psychol 2021 Aug 10;76(5):744-754. [doi: 10.1037/amp0000717] [Medline: 32772540]

10. Pepe E, Bajardi P, Gauvin L, Privitera F, Lake B, Cattuto C, et al. COVID-19 outbreak response, a dataset to assess mobility changes in Italy following national lockdown. Sci Data 2020 Jul 08;7(1):230 [FREE Full text] [doi: 10.1038/s41597-020-00575-2] [Medline: 32641758]

11. Kishore N, Kiang MV, Engø-Monsen K, Vembar N, Schroeder A, Balsari S, et al. Measuring mobility to monitor travel and physical distancing interventions: a common framework for mobile phone data analysis. Lancet Digital Health 2020 Nov;2(11):e622-e628 [FREE Full text] [doi: 10.1016/S2589-7500(20)30193-X]

12. Oliver N, Lepri B, Sterly H, Lambiotte R, Deletaille S, De Nadai M, et al. Mobile phone data for informing public health actions across the COVID-19 pandemic life cycle. Sci Adv 2020 Jun 27;6(23):eabc0764 [FREE Full text] [doi: 10.1126/sciadv.abc0764] [Medline: 32548274]

13. Kraemer MUG, Yang C, Gutierrez B, Wu C, Klein B, Pigott DM, Open COVID-19 Data Working Group, et al. The effect of human mobility and control measures on the COVID-19 epidemic in China. Science 2020 May 01;368(6490):493-497 [FREE Full text] [doi: 10.1126/science.abb4218] [Medline: 32213647]

14. Ruktanonchai NW, Floyd JR, Lai S, Ruktanonchai CW, Sadilek A, Rente-Lourenco P, et al. Assessing the impact of coordinated COVID-19 exit strategies across Europe. Science 2020 Sep 18;369(6510):1465-1470 [FREE Full text] [doi: 10.1126/science.abc5096] [Medline: 32680881]

15. Gao S, Rao J, Kang Y, Liang Y, Kruse J, Dopfer D, et al. Association of mobile phone location data indications of travel and stay-at-home mandates with COVID-19 infection rates in the US. JAMA Netw Open 2020 Sep 01;3(9):e2020485 [FREE Full text] [doi: 10.1001/jamanetworkopen.2020.20485] [Medline: 32897373]

16. Floyd D, Prentice-dunn S, Rogers R. A meta-analysis of research on protection motivation theory. J Appl Social Pyschol 2000 Feb;30(2):407-429 [FREE Full text] [doi: 10.1111/j.1559-1816.2000.tb02323.x]

17. Brug J, Aro AR, Richardus JH. Risk perceptions and behaviour: towards pandemic control of emerging infectious diseases; international research on risk perception in the control of emerging infectious diseases. Int J Behav Med 2009 Jan 6;16(1):3-6 [FREE Full text] [doi: 10.1007/s12529-008-9000-x] [Medline: 19127440]

18. Rogers RW. A protection motivation theory of fear appeals and attitude change1. J Psychol 1975 Sep 02;91(1):93-114. [doi: 10.1080/00223980.1975.9915803] [Medline: 28136248]

19. Pedro SA, Ndjomatchoua FT, Jentsch P, Tchuenche JM, Anand M, Bauch CT. Conditions for a second wave of COVID-19 due to interactions between disease dynamics and social processes. Front Phys 2020 Oct 9;8:574514. [doi: 10.3389/fphy.2020.574514]

20. Panovska-Griffiths J, Kerr CC, Stuart RM, Mistry D, Klein DJ, Viner RM, et al. Determining the optimal strategy for reopening schools, the impact of test and trace interventions, and the risk of occurrence of a second COVID-19 epidemic wave in the UK: a modelling study. Lancet Child Adolesc Health 2020 Nov;4(11):817-827. [doi: 10.1016/s2352-4642(20)30250-9]

21. Yamamoto T, Uchiumi C, Suzuki N, Yoshimoto J, Murillo-Rodriguez E. The psychological impact of 'mild lockdown' in Japan during the COVID-19 pandemic: a nationwide survey under a declared state of emergency. Int J Environ Res Public Health 2020 Dec 15;17(24):9382 [FREE Full text] [doi: 10.3390/ijerph17249382] [Medline: 33333893]

22. Masayuki T, Tomohiro N, Motonari K. "Mobile spatial statistics" supporting development of society and industry: population estimation technology using mobile network statistical data and applications. NTT DOCOMO Tech J 2013;14(3):10-15 [FREE Full text]

23. Statistics Bureau of Japan. Economic Census for Business Frame. URL: https://www.stat.go.jp/english/data/e-census/index.html [accessed 2022-03-18]

24. Brodersen KH, Gallusser F, Koehler J, Remy N, Scott SL. Inferring causal impact using Bayesian structural time-series models. Ann Appl Stat 2015 Mar 1;9(1):247-274. [doi: 10.1214/14-aoas788]

25. Google. Google COVID-19 Community Mobility Reports. URL: https://www.google.com/covid19/mobility/ [accessed 2022-03-18]

26. Rypdal K, Bianchi FM, Rypdal M. Intervention fatigue is the primary cause of strong secondary waves in the COVID-19 pandemic. Int J Environ Res Public Health 2020 Dec 21;17(24):9592 [FREE Full text] [doi: 10.3390/ijerph17249592] [Medline: 33371489]

27. Yabe T, Tsubouchi K, Fujiwara N, Wada T, Sekimoto Y, Ukkusuri SV. Non-compulsory measures sufficiently reduced human mobility in Tokyo during the COVID-19 epidemic. Sci Rep 2020 Oct 22;10(1):18053 [FREE Full text] [doi: 10.1038/s41598-020-75033-5] [Medline: 33093497]

Abbreviations

CCF: cross-correlation function
BSTS: Bayesian structural time series
RQ: research question

XSL·FO

RenderX

<u>Original Paper</u>

Disease-Course Adapting Machine Learning Prognostication Models in Elderly Patients Critically Ill With COVID-19: Multicenter Cohort Study With External Validation

Christian Jung[1], MD, PhD[‡]; Behrooz Mamandipoor[2], BSc; Jesper Fjølner[3], MD; Raphael Romano Bruno[1], MD[‡]; Bernhard Wernly[4], MD, PhD; Antonio Artigas[5], MD; Bernardo Bollen Pinto[6]; Joerg C Schefold[7]; Georg Wolff[1], MD; Malte Kelm[1], MD; Michael Beil[8]; Sigal Sviri[8], MHA; Peter V van Heerden[9], MD, PhD; Wojciech Szczeklik[10]; Miroslaw Czuczwar[11]; Muhammed Elhadi[12]; Michael Joannidis[13]; Sandra Oeyen[14]; Tilemachos Zafeiridis[15†]; Brian Marsh[16]; Finn H Andersen[17,18]; Rui Moreno[19,20], MD, PhD; Maurizio Cecconi[21]; Susannah Leaver[22]; Dylan W De Lange[23]; Bertrand Guidet[24,25]; Hans Flaatten[26,27]; Venet Osmani[2], PhD

[1]Division of Cardiology, Pulmonology and Vascular Medicine, Medical Faculty, Heinrich-Heine-University Duesseldorf, University Hospital Duesseldorf, Duesseldorf, Germany

[2]Fondazione Bruno Kessler Research Institute, Trento, Italy

[3]Department of Intensive Care, Aarhus University Hospital, Aarhus, Denmark

[4]Department of Anaesthesiology, Paracelsus Medical University, Salzburg, Austria

[5]Department of Intensive Care Medicine, CIBER Enfermedades Respiratorias, Corporacion Sanitaria Universitaria Parc Tauli, Autonomous University of Barcelona, Sabadell, Spain

[6]Department of Acute Medicine, Geneva University Hospitals, Geneva, Switzerland

[7]Department of Intensive Care Medicine, Inselspital, Universitätsspital, University of Bern, Bern, Switzerland

[8]Department of Medical Intensive Care, Hadassah University Medical Center, Faculty of Medicine, Hebrew University of Jerusalem, Jerusalem, Israel

[9]Department of Anesthesia, Intensive Care and Pain Medicine, Hadassah Medical Center and Faculty of Medicine, Hebrew University of Jerusalem, Jerusalem, Israel

[10]Center for Intensive Care and Perioperative Medicine, Jagiellonian University Medical College, Krakow, Poland

[11]Second Department of Anesthesiology and Intensive Care, Medical University of Lublin, Lublin, Poland

[12]Faculty of Medicine, University of Tripoli, Tripoli, Libyan Arab Jamahiriya

[13]Division of Intensive Care and Emergency Medicine, Department of Internal Medicine, Medical University Innsbruck, Innsbruck, Austria

[14]Department of Intensive Care 1K12IC, Ghent University Hospital, Ghent, Belgium

[15]Intensive Care Unit, General Hospital of Larissa, Larissa, Greece

[16]Mater Misericordiae University Hospital, Dublin, Ireland

[17]Department of Anaesthesia and Intensive Care, Ålesund Hospital, Alesund, Norway

[18]Department of Circulation and Medical Imaging, Norwegian University of Science and Technology, Trondheim, Norway

[19]Hospital de São José, Centro Hospitalar Universitário de Lisboa Central, Lisbon, Portugal

[20]Faculdade de Ciências Médicas de Lisboa, Nova Medical School - Faculdade de Ciências Médicas, Universidade da Beira Interior, Lisbon, Portugal

[21]Department of Anaesthesia, IRCCS Instituto Clínico Humanitas, Humanitas University, Milan, Italy

[22]General Intensive Care, St George's University Hospitals, NHS Foundation Trust, London, United Kingdom

[23]Department of Intensive Care Medicine, University Medical Center, Utrecht University, Utrecht, Belgium

[24]Épidémiologie Hospitalière Qualité et Organisation des Soins, Institut Pierre Louis d'Epidémiologie et de Santé Publique, Sorbonne Universités, UPMC Univ Paris 06, INSERM, UMR_S 1136, Paris, France

[25]Service de Réanimation Médicale, Assistance Publique-Hôpitaux de Paris, Hôpital Saint-Antoine, Paris, France

[26]Department of Clinical Medicine, University of Bergen, Bergen, Norway

[27]Department of Anesthesia and Intensive Care, Haukeland University Hospital, Bergen, Norway

[†]deceased[‡]COVIP Study Group

Corresponding Author:
Christian Jung, MD, PhD
Division of Cardiology, Pulmonology and Vascular Medicine
Medical Faculty, Heinrich-Heine-University Duesseldorf
University Hospital Duesseldorf
Moorenstraße 5

Duesseldorf, 40225
Germany
Phone: 49 2118118800
Fax: 49 211 81 19520
Email: christian.jung@med.uni-duesseldorf.de

Abstract

Background: The COVID-19 pandemic caused by SARS-CoV-2 is challenging health care systems globally. The disease disproportionately affects the elderly population, both in terms of disease severity and mortality risk.

Objective: The aim of this study was to evaluate machine learning–based prognostication models for critically ill elderly COVID-19 patients, which dynamically incorporated multifaceted clinical information on evolution of the disease.

Methods: This multicenter cohort study (COVIP study) obtained patient data from 151 intensive care units (ICUs) from 26 countries. Different models based on the Sequential Organ Failure Assessment (SOFA) score, logistic regression (LR), random forest (RF), and extreme gradient boosting (XGB) were derived as baseline models that included admission variables only. We subsequently included clinical events and time-to-event as additional variables to derive the final models using the same algorithms and compared their performance with that of the baseline group. Furthermore, we derived baseline and final models on a European patient cohort, which were externally validated on a non-European cohort that included Asian, African, and US patients.

Results: In total, 1432 elderly (≥70 years old) COVID-19–positive patients admitted to an ICU were included for analysis. Of these, 809 (56.49%) patients survived up to 30 days after admission. The average length of stay was 21.6 (SD 18.2) days. Final models that incorporated clinical events and time-to-event information provided superior performance (area under the receiver operating characteristic curve of 0.81; 95% CI 0.804-0.811), with respect to both the baseline models that used admission variables only and conventional ICU prediction models (SOFA score, $P<.001$). The average precision increased from 0.65 (95% CI 0.650-0.655) to 0.77 (95% CI 0.759-0.770).

Conclusions: Integrating important clinical events and time-to-event information led to a superior accuracy of 30-day mortality prediction compared with models based on the admission information and conventional ICU prediction models. This study shows that machine-learning models provide additional information and may support complex decision-making in critically ill elderly COVID-19 patients.

Trial Registration: ClinicalTrials.gov NCT04321265; https://clinicaltrials.gov/ct2/show/NCT04321265

Introduction

The COVID-19 pandemic caused by SARS-CoV-2 is continuing to challenge health care systems globally [1]. The disease disproportionately affects the elderly population, both in terms of disease severity and mortality risk [2]. In many countries, intensive care unit (ICU) capacity was increased during the pandemic to meet demand. In addition, novel treatment modalities were introduced [3]. A key challenge in clinical outcome prediction in a dynamic disease is that the response to a given treatment varies considerably from patient to patient, especially in the elderly population [4]. Baseline data alone are inadequate to predict prognosis with sufficient accuracy for an individual patient, as they cannot capture the dynamic nature of the underlying critical illness [5]. It is well established that various factors provide prognostic information that should be taken into consideration [6]. More elaborate methods are thus urgently needed for both sophisticated and concise risk stratification of severely affected individual ICU patients [7]. Biomarkers, frailty, and severity scores are validated in elderly critically ill patients [8-11]. However, all of these have important

limitations as they do not reflect the dynamics of the underlying disease pathophysiology, and as a result have limited prognostic power. Ultimately, it remains up to the physician to integrate all baseline data, the changing course of the disease, and subjective experience into a clinical decision [12]. However, physicians do not assess dynamically evolving processes perfectly, as they are influenced by numerous factors, including fatigue and other human factors, resulting in less objective and reproducible decision-making [13]. This aspect is especially relevant for new diseases such as COVID-19, where physician experience is lacking.

Therefore, a supportive prognostication model that can integrate baseline data with complex, dynamic processes in an objective manner is necessary. Machine learning (ML) algorithms could be used to address this need, as some have successfully been evaluated in clinical settings such as in cardiovascular intensive care [14]. Wernly et al [9] retrospectively analyzed arterial blood gas data from septic intensive care patients from a multicenter electronic ICU database as well as from a single-center MIMIC-III (Medical Information Mart for Intensive Care) data set to predict 96-hour mortality.

Izquierdo et al [15] combined classical epidemiological methods, natural language processing, and ML to examine the electronic health records of 10,504 patients with COVID-19. According to their analysis, the combination of easily obtainable clinical variables such as age, fever, and tachypnea predicted which patients would require ICU admission [15]. The observational study by Bolourani et al [16] had a similar aim. They used clinical and laboratory data commonly collected in the emergency department to train and validate three predictive models (two based on extreme gradient boosting [XGB] and one that used logistic regression [LR]) with cross-hospital validation. The XGB model had the highest mean accuracy to predict 48-hour respiratory failure [16]. Aktar et al [17] used ML to distinguish between healthy people and those with COVID-19 and subsequently to predict COVID-19 severity. They used decision tree, random forest (RF), variants of gradient boosting machine, support vector machine, k-nearest neighbor, and deep learning methods for blood samples. The developed analytical methods evidenced accuracy and precision scores >90% for disease severity prediction. To avoid locally aggregating raw clinical data across multiple institutions, Vaid et al [18] evaluated a federated learning ML technique using electronic health records from 5 hospitals. In brief, they used LR with L1 regularization/least absolute shrinkage and selection operator, and multilayer perceptron models that were trained using local data at each study site. The federated models outperformed the local models with regard to their accuracy in predicting the mortality in hospitalized patients with COVID-19 within 7 days. In a smaller study, Domínguez-Olmedo et al [19] selected 32 predictor laboratory features in 1823 patients with confirmed COVID-19 for an XGB algorithm. Similar to the other studies, using laboratory parameters resulted in excellent outcome prediction. Subudhi et al [20] used ensemble-based ML models to identify C-reactive protein, lactate dehydrogenase, and oxygen saturation as the most important factors for predicting ICU admission, with estimated glomerular filtration rate <60 mL/min/1.73 m^2, and neutrophil and lymphocyte percentages as the important factors for predicting mortality.

A recent systematic review by Syeda et al [21] identified more than 400 articles that investigated the role of ML in the field of COVID-19. For example, Pan et al [22] studied 123 ICU patients and identified eight important risk factors with high recognition ability using an XGB model. A similar approach was used by Kim et al [23], who established an XGB model in 4787 patients admitted to a hospital due to COVID-19. Furthermore, Burian et al [24] estimated the need for intensive care treatment in 65 patients with confirmed COVID-19, and Shahsikumar et al [25] investigated the performance of an algorithm to predict the need for mechanical ventilation on 402 patients with COVID-19, using cohorts with a wide age range (48 to 74 years).

Patients who are very old represent the most vulnerable intensive care subgroup [26]. However, to date, there are no studies investigating the role of ML models in this specific subgroup exclusively. To address this lack of evidence, the aim of this study was to evaluate whether ML models can reliably improve mortality prognostication in critically ill elderly patients with COVID-19 based on clinical baseline information, biomarkers, accumulating events, and time-to-event information during the disease course.

Methods

Study Design

This was a retrospective analysis that included data from 1432 patients in a prospective multicenter study. The primary outcome was 30-day mortality. We also used the 3-month outcome to ensure consistency of the primary outcome and allay concerns of censoring bias [27]. We derived two groups of models: baseline and final models. Baseline models were derived using admission variables only, whereas the final model group incorporated clinical events such as catecholamine therapy, renal replacement therapy, noninvasive ventilation, invasive ventilation, prone position, and tracheostomy, in addition to the baseline variables. We evaluated both model groups using stratified 3-fold cross-validation to mitigate the variability of a single derivation–validation random split. Furthermore, we derived baseline and final models on an EU patient cohort and externally validated them on a non-EU cohort that included Asian, African, and US patients.

Clinical Data Sources and Study Population

Patient data were obtained from 151 ICUs across 26 independent countries, including European ICUs, and from ICUs in Asia, Africa, and the United States as part of the multinational COVIP trial (NCT04321265). This study was conducted in line with the European Union General Data Privacy Regulation directive. As in previous successful studies [6,26,28], national coordinators recruited the ICUs, coordinated national and local ethical permissions, and supervised patient recruitment at the national level. In the COVIP studies, ethical approval was obligatory for study participation. The electronic case report form (eCRF) and database were hosted on a secure server in Aarhus University, Denmark. Data from 1432 elderly (aged 70 years and above) COVID-19–positive patients admitted to a participating ICU between February 4 and May 26, 2020, were recorded. The study protocol is available from the COVIP study website [29]. Patients were followed up until hospital discharge and survival at 3 months using telephone interviews.

Ethical Considerations

The primary competent ethics committee was the Ethics Committee of the University of Duesseldorf, Germany. Institutional research ethics board approval was obtained from each study site. This was a prerequisite for participation in the study. All methods were carried out in accordance with relevant guidelines and regulations. All experimental protocols were approved by the local institutional and/or licensing committees. Informed consent was obtained from all subjects if not omitted by the ethics vote. The studies were all observational; no examinations (eg, blood sampling) or tissue sampling took place.

Study Data

Demographic data included age, gender, weight, height, and BMI. Furthermore, information on admission characteristics prior to ICU hospitalization, duration of hospital stay, day of symptom onset, and comorbidities were available. Preexisting

comorbidities were recorded in the eCRF: diabetes, ischemic heart disease, renal insufficiency, arterial hypertension, pulmonary comorbidity, and chronic heart failure.

During the ICU stay, data on bacterial coinfection were noted, in addition to Sequential Organ Failure Assessment (SOFA) subscores (respiratory, cardiovascular, hepatic, coagulation, renal, and neurological systems). Laboratory values included partial oxygen pressure and the fraction of inspired oxygen (FiO2), and their ratio. Six clinical events of interest (catecholamine therapy, renal replacement therapy, noninvasive and invasive ventilation, prone position, and tracheostomy) were recorded along with the time the event occurred.

Model Derivation and Validation

We derived models based on XGB [30], RF [31], and LR [32]. As the best-performing model, the XGB algorithm provides robust prediction results using a method where new models are added to correct the errors made by existing models. Models are added sequentially and the combination of many models in the XGB model accommodates nonlinearity between input variables [30]. Hyperparameter tuning was performed by an exhaustive grid search directed toward maximizing the F1-score metric. Three-fold cross-validation was performed inside each grid option, and the optimal hyperparameter set was chosen based on the model in the grid search with the highest F1 score. Hyperparameters of the final model of the XGB are listed in Multimedia Appendix 1. To generate confidence intervals for the baseline and the final models, 3-fold cross-validation was performed with 20-times repetition with a randomly generated seed. To compare the performance of the XGB model, we also derived and validated two more predictive models based on LR and RF. This decision was driven by the fact that LR is typically considered a baseline algorithm, and RF has been previously used in other research with COVID-19 data [33]. Both LR and RF were optimized by an exhaustive grid search, similar to the XGB method.

To address noise and outliers in the data, we defined a clinically valid interval for each variable, and the values out of the valid scope were considered as missing values. For all models, the issue of missing values was addressed by removing variables with >90% missing values. We then used the median and zero to impute the missing data in the remaining continuous and categorical variables, respectively. All analyses were carried out using open-source software based on Python 3.6.8 with scikit-learn version 0.23.2.

Experimental Evaluation

Performance evaluation of the models was based on 3-fold, stratified cross-validation with 20 repetitions using the area under the receiver operating characteristic curve (AUC; see step 3 in Figure 1) as well as area under the precision-recall curve (PRC), also known as average precision [34].

Figure 1. Graphical methods. (1) Study design, from admission to derivation and validation of baseline setup. (2) Derivation and validation of six models incorporating clinical events individually.Performance of individual models is shown in Multimedia Appendix 2-5. (3) Derivation of the final model, including baseline variables as well as clinical events. (4) Evaluation of the final model in predicting 30-day outcomes. SOFA: Sequential Organ Failure Assessment; ICU: intensive care unit.
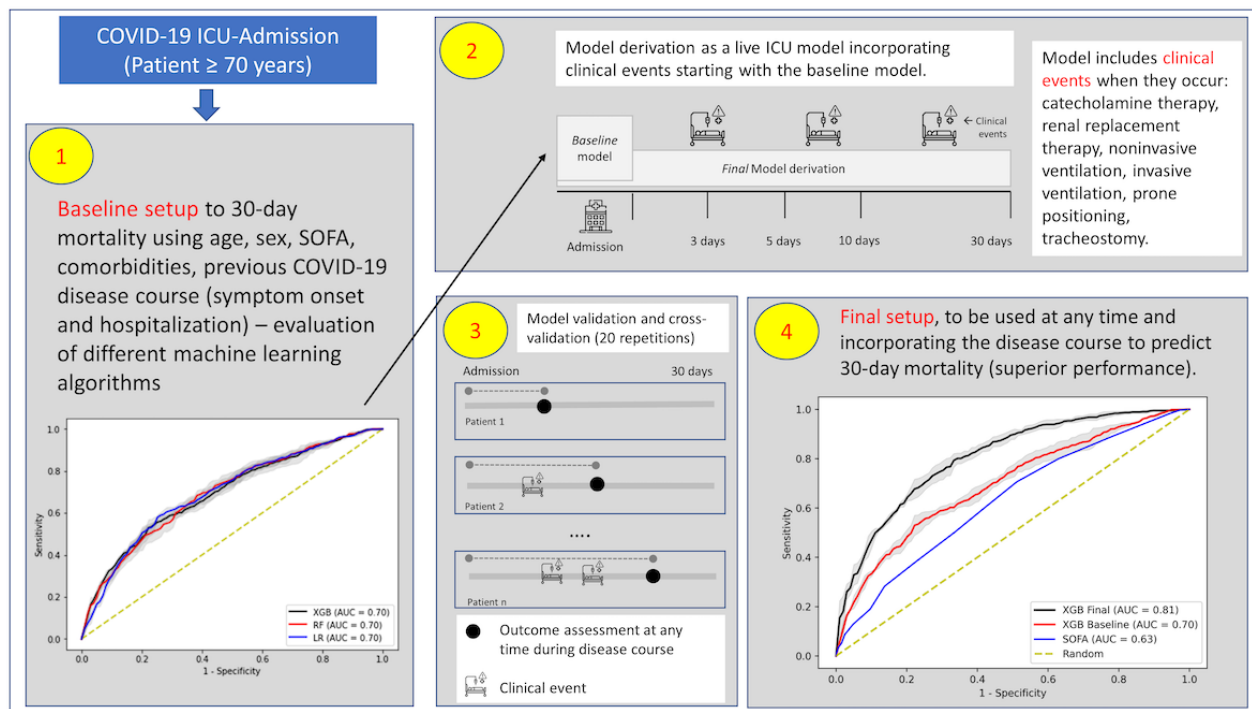
The PRC shows the relationship between the positive predictive value (precision) and sensitivity (recall), measuring the performance of the model in correctly predicting mortality in patients with a high probability of dying. The area under the PRC is typically more informative than the AUC in the presence of imbalanced outcomes [34]. Additional performance metrics are detailed in Multimedia Appendix 2-5, including the positive predictive value (PPV), negative predictive value, F1 score (the balance between PPV and sensitivity), Matthews correlation coefficient (used to measure the quality of classification between

algorithms), and Brier score. Calibration quality was evaluated using Brier scores, where a lower score indicates a higher calibration quality, and we also present calibration plots (also known as reliability curves). The models were compared based on their AUC and PRC performance metrics for both the baseline data as well as the final models incorporating clinical events.

Model Interpretation

We evaluated the ranking of variables that contributed toward the model description using shapely additive explanation (SHAP) scores. SHAP scores are a game-theoretic approach to model interpretability; they provide explanations of global model structures based on combinations of several local explanations for each prediction [35]. To interpret and rank the significance of input variables toward the final prediction of the model, mean absolute SHAP values were calculated for each variable across all observations in both the baseline model and the final model based on XGB. We also plotted SHAP interaction values that capture the contribution of pairwise interactions between unique features to model prediction. To improve interpretability, especially in terms of the impact of clinical events, we defined a clinically meaningful day interval (0-3, 3-5, 5-10, and 10-30 days), and added a variable for each clinical event based on when the clinical event occurred; for example, "Tracheostomy-10-30" indicates that a tracheostomy was performed within the 10-30–day period. This allowed us to evaluate not only the importance of clinical events but also the time-to-event information. Naturally, these variables were only available in the final model.

Results

Study Population

Out of the total 1432 patients in the COVIP cohort, 809 (56.49%) patients survived up to 30 days after admission, with an average length of stay of 21.6 (SD 18.2) days. Patient baseline characteristics are given in Table 1, with distribution of mortality and length of stay detailed in Multimedia Appendix 6.

Table 1. Demographic characteristics, vital signs, and clinical events of patient cohorts (N=1432).

Variables	Alive at 30 days (n=809)	Dead at 30 days (n=623)	*P* value
Sex (male), n (%)	587 (72.6%)	463 (74.6%)	.18
Age (years), mean (SD)	75.0 (4.2)	76.5 (4.8)	<.001
Weight (kg), mean (SD)	81.3 (14.7)	81.0 (14.8)	.42
Height (cm), mean (SD)	169.7 (10.7)	169.8 (10.5)	.06
BMI (kg/m²), mean (SD)	28.5 (6.5)	28.4 (5.7)	.02
Hospital stay prior to ICU[a] admission (days), mean (SD)	3.8 (5.7)	3.5 (6.3)	.002
Symptoms prior to hospital admission (days), mean (SD)	7.2 (5.2)	6.6 (4.5)	.10
PaO2[b] (mmHg), mean (SD)	87.3 (44.2)	84.3 (57.5)	.003
FiO2[c] (%), mean (SD)	62.3 (31.0)	73.0 (24.0)	<.001
SOFA[d] score (points), mean (SD)	5.2 (3.0)	6.7 (3.4)	<.001
ICU treatment and outcome			
Mechanical ventilation, n (%)	561 (69.3)	510 (81.9)	<.001
Vasopressors, n (%)	525 (64.9)	515 (82.7)	<.001
Prone positioning, n (%)	309 (38.2)	279 (44.8)	.10
Tracheostomy, n (%)	227 (28.1)	64 (10.3)	<.001
Noninvasive ventilation, n (%)	169 (20.9)	119 (19.1)	.32
Renal replacement therapy, n (%)	121 (15.0)	119 (19.1)	.01
Length of ICU stay (days), mean (SD)	21.6 (18.2)	10.6 (7.6)	<.001
Preexisting comorbidities, n (%)			
Diabetes mellitus	268 (33.1)	240 (38.5)	.01
Ischemic heart disease	151 (18.7)	152 (24.4)	.007
Chronic renal insufficiency	91 (11.2)	130 (20.9)	<.001
Arterial hypertension	527 (65.1)	431 (69.2)	.03
Pulmonary disease	175 (21.6)	145 (23.3)	.07
Chronic heart failure	98 (12.1)	103 (16.5)	.01

[a]ICU: intensive care unit.

[b]PaO2: partial oxygen pressure.

[c]FiO2: fraction of inspired oxygen.

[d]SOFA: Sequential Organ Failure Assessment.

Model Derivation and Validation

We evaluated the performance of *baseline setup* risk prognostication that included baseline variables only (see step 1 in Figure 1) and the *final* setup, which—in addition to baseline variables—included six key clinical events that occurred during the disease course and their time-to-event information: catecholamine therapy, renal replacement therapy, noninvasive ventilation, invasive ventilation, prone positioning, and tracheostomy (step 2 in Figure 1). The final set of selected variables is shown in Table 1. Furthermore, the baseline and the final setup were used to derive models on the EU cohort of

patients that were then externally evaluated using a non-EU cohort composed of Asian, African, and US patients.

Three risk prognostication models were derived from ML-based algorithms: LR and, for comparison, RF and XGB algorithms, as outlined in the Methods section [30,31].

The XGB algorithm achieved the numerically highest increase in discrimination performance from the *baseline setup* (AUC 0.70, 95% CI 0.692-0.701) to the *final setup* (AUC 0.81, 95% CI 0.804-0.811); average precision increased from 0.65 (95% CI 0.650-0.655) to 0.77 (95% CI 0.759-0.770) (Figure 2).

Figure 2. Performance of the baseline model (top) and improved performance in the final model (bottom) in response to clinical events with respect to the area under the receiver operating characteristic (ROC) curve (AUC) and area under the precision-recall curve (PRC). The PRC shows the relationship between the positive predictive value (precision) and sensitivity (recall) at all thresholds. XGB: extreme gradient boosting; RF: random forest; LR: logistic regression; SOFA: Sequential Organ Failure Assessment.
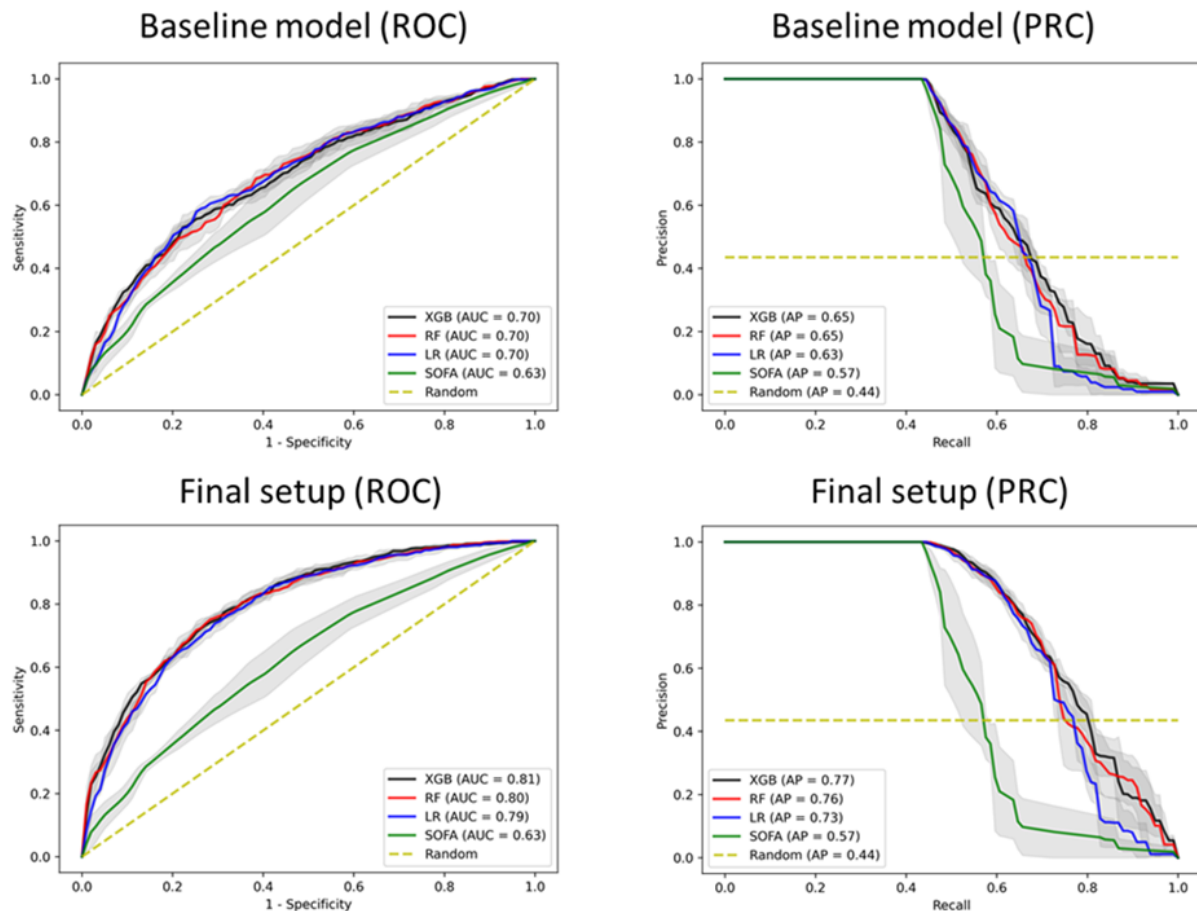
The LR (AUC 0.79, 95% CI 0.788-0.796) and RF (AUC 0.80, 95% CI 0.798-0.805) algorithms showed similar performance in the *baseline model* and improvement in the *final model*, comparable to XGB performance (see step 4 in Figure 1). The final XGB model provided superior performance compared to both the baseline model and SOFA score (both *P*<.001).

Experimental Evaluation

In the external validation of the EU patient cohort, all three models achieved similar performance in the baseline and the final setup with an AUC of 0.82 and 0.86, respectively, when

evaluated on predicting the mortality of non-EU patients (Figure 3). One explanation for this performance on the external validation cohort might be that the patients in the non-EU cohort tended to gravitate toward two opposing health states of either being quite stable or very sick, making it easier for the model to discriminate between the two outcomes. To investigate this further, we plotted the distribution of the variable that had the highest impact on outcome prediction (FiO2) based on SHAP analysis (see Figure 4). As shown in Multimedia Appendix 7, the distribution for both outcomes was significantly skewed toward 21% for survivors and toward 100% for nonsurvivors.

Figure 3. Performance of the final model derived using the EU patient cohort and externally validated on a non-EU patient cohort, comprising Asian, African, and US patients. Model performance is measured using area under the receiver operating characteristic (ROC) curve (AUC) and area under the precision-recall curve (PRC). XGB: extreme gradient boosting; RF: random forest; LR: logistic regression.
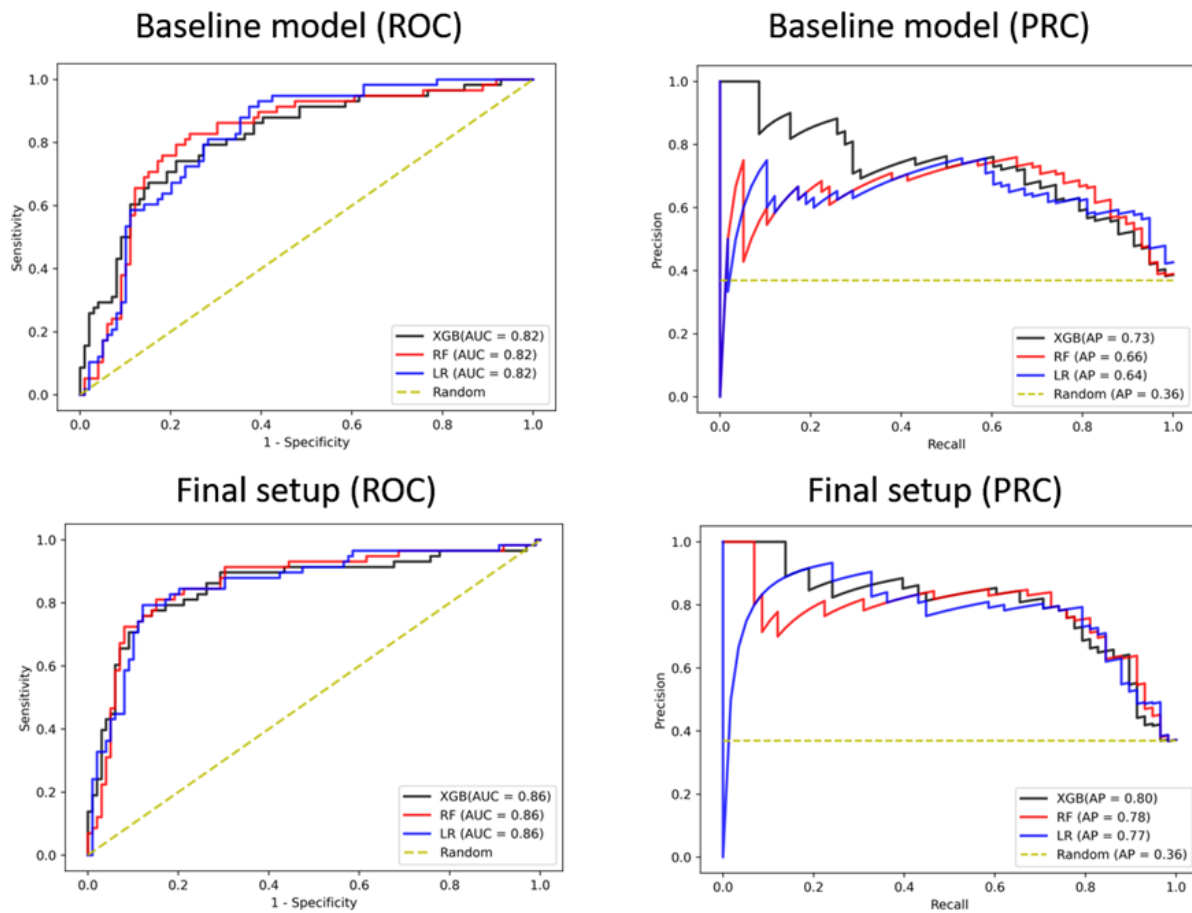
We also assessed the calibration of each model to ensure that the distribution of predicted outcomes matches the distribution of observed outcomes in our patient cohort. Baseline and final models were, in general, well calibrated (Figure 5), matching the estimated risk of outcome with observed risk. The final setup for each algorithm was better calibrated (Brier score of 0.17) with respect to the baseline setup (Brier score 0.22). Full details of Brier scores for each algorithm are detailed in Multimedia Appendix 1.

Figure 4. Ranking of input variables of the final setup derived from the extreme gradient boost algorithm, using the shapely additive explanation (SHAP) method.
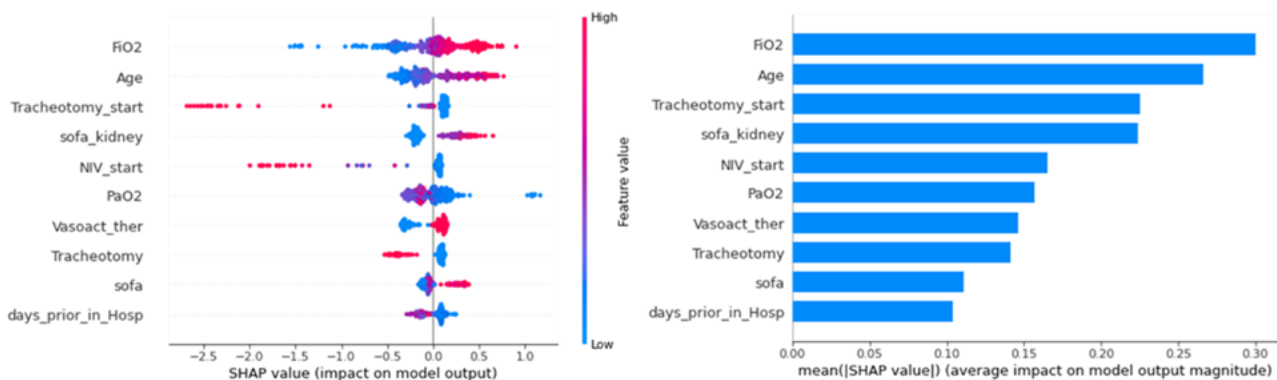
Figure 5. Calibration curves for each model and individual algorithms used to derive the model. XGB, extreme gradient boosting; RF: random forest; LR: logistic regression.
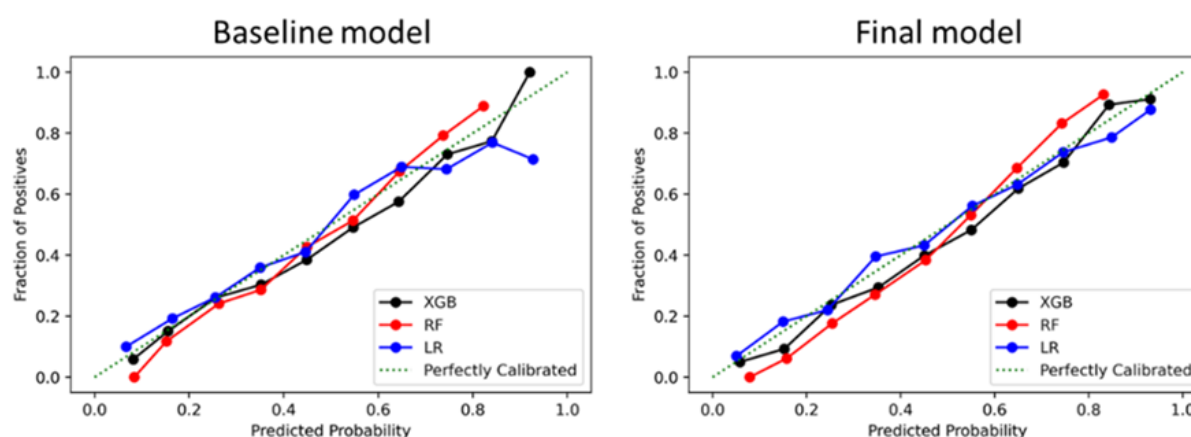
Model Interpretation

The SHAP method was used to perform interpretability analysis, which explains model output by computing the contribution of each variable to the prediction. Among others, the SHAP method was applied on the best-performing model (XGB), where the FiO2, age, and tracheostomy had the highest impact on outcome prediction (Figure 4 and Multimedia Appendix 7).

We also report the model interpretability analysis for the RF- and LR-based models in Multimedia Appendix 8 and 9, respectively. The top three variables remained common between XGB and RF, whereas for LR, only tracheostomy appeared in the top three, with the other two high-ranking variables being weight and BMI.

Discussion

Principal Findings and Comparison With Related Studies

This study demonstrates that individual prognostication accuracy based on patient baseline characteristics can be considerably improved with ML algorithms that incorporate occurrence and time-to-event information of clinical events along the course of a disease such as COVID-19 in elderly, critically ill patients. These results align with many previous studies that investigated ML approaches in patients suffering from COVID-19. The major difference between this COVIP study and others published previously lies in its focus on the especially vulnerable subgroup of very old intensive care patients [21]. The second important difference is that the current approach includes the risk for clinical events such as tracheostomy.

Subudhi et al [20] compared the ability of 18 different ML algorithms to predict the rate of admission and mortality of patients suffering from COVID-19. In their analysis, ensemble-based models were superior to other algorithms (including LR and XGB). Specific laboratory values and oxygen saturation were the most important factors for ICU admission, whereas impaired kidney function and differential blood count best predicted mortality [20]. However, this previous study primarily used data from patients, of all ages, presenting to the emergency room.

Domínguez-Olmedo et al [19] used data from 1823 patients with confirmed COVID-19 and established an XGB model. Their model found lactate dehydrogenase activity, C-reactive protein level, neutrophil count, and urea level to be the most important variables, reaching an AUC of 0.93 (95% CI 0.89-0.98) for sensitivity and 0.91 (95% CI 0.86-0.96) for specificity.

Pan et al [22] used data from 123 patients with COVID-19 admitted to an ICU to construct an XGB model, and identified eight factors (albumin level, creatinine, eosinophil percentage, lactate dehydrogenase, lymphocyte percentage, neutrophil percentage, prothrombin time, and total bilirubin) that were predictive for ICU mortality.

Vaid et al [18] utilized a different approach based on federated learning of electronic health records from five different hospitals, providing robust predictive models without compromising patient privacy.

Other studies focused primarily on peripheral blood samples. Aktar et al [17] developed ML and deep learning algorithms to predict the disease severity. Similarly, Kim et al [23] established an XGB model in 4787 hospital-admitted patients to predict their intensive care treatment requirements. Their model was significantly superior to the established CURB-65 (confusion, urea, respiratory rate, blood pressure) score.

Applications

Immediate clinical applications are conceivable, especially given the limited number of ICU beds available. Our models may be used in several ways: ML could be used before ICU admission to offer objective support for complex allocation decisions. However, ML algorithms would mainly access data at presentation and few dynamic parameters, limiting the predictive power. ML algorithms could also be used in the context of time-limited trials (TLTs), which are common clinical practice in ICUs in some countries. This may be particularly helpful in patients for whom realistic therapeutic goals/outcomes are unclear at presentation. These patients could be admitted to the ICU under the premise of gaining more information about the patient and the initial response to treatment. This additional information could then be evaluated using ML algorithms [36] as already shown in patients with sepsis [9]. The ideal temporal

combination of a TLT and ML should be the subject of future, prospective studies [36,37].

In terms of practical applications, ML algorithms provide a potential strategy to improve decision confidence and predictive power over time. They are applicable at various time points during the disease course, predicting outcomes in a continuous manner. This approach is especially applicable when considering that the model was well calibrated in estimating outcomes. However, evaluation of the model with a diverse patient population would provide further evidence of its clinical applicability.

Clinical evaluations such as assessment of wakefulness, mobility, responsiveness, and independence are subjective and subject to interrater variability. Therefore, advances in digital technologies may support but not replace physicians' skills. ML can support physicians, especially in estimations on prognosis and achievement of therapy goals. Importantly, ethical problems become evident when ML is involved in matters of life and death [38], and it must be emphasized that ML should only support and aid medical decision-making. Our data show that dedicated modern algorithms can incrementally improve certainty during TLTs in elderly patients with COVID-19, and generalize well in an external patient cohort. These tools can enhance our ability to improve guidance of treatment and optimally allocate ICU resources. However, such a strategy can only be viewed as complementary to clinical judgment and individual treatment goals, and form part of a holistic patient assessment.

Limitations

This study has some methodological limitations in common with the other COVIP studies [11,26,39-42]. COVIP did not contain a control group of younger COVID-19 patients for comparison or a comparable age cohort of patients who were not or could not be admitted to the ICU. In addition, the COVIP database does not include information on pre-ICU care and triage decisions. These treatment limitations might also affect the care of older ICU patients [43]. Furthermore, COVIP recruited patients in 26 countries, and thus the participating countries varied widely in their care structure, resulting in considerable heterogeneity in treatments given.

Conclusion

This study demonstrates that, in the particularly vulnerable subgroup of very old intensive care patients suffering from COVID-19, individual prognostication accuracy based on patient baseline characteristics can be improved with ML algorithms. These algorithms capture the dynamic course of the disease by including the occurrence and time-to-event information of clinical events, and thus reflect both disease severity and the need for intensive care treatment.

Authors' Contributions

BW, BM, JF, RB, VO, and CJ analyzed the data and wrote the first draft of the manuscript. AA, BBP, JCS, and GW contributed to the statistical analysis and improved the paper. MK, MB, SS, PVH, WS, MC, ME, MJ, SO, TZ, BM, FA, RM, MC, SL, DWDL, BG, and HF gave guidance and improved the paper. All authors read and approved the final manuscript.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Hyperparameters for each algorithm found through an exhaustive grid search.
[DOCX File , 14 KB - medinform_v10i3e32949_app1.docx]

Multimedia Appendix 2
Performance of the baseline model in terms of various performance metrics and 95% CIs: logistic regression (LR), random forest (RF), extreme gradient boosting (XGB).
[DOCX File , 14 KB - medinform_v10i3e32949_app2.docx]

Multimedia Appendix 3
Performance of the final model in terms of various performance metrics and 95% CIs: logistic regression (LR), random forest (RF), extreme gradient boosting (XGB).
[DOCX File , 14 KB - medinform_v10i3e32949_app3.docx]

Multimedia Appendix 4

Performance of the baseline model derived using the EU patient cohort and validated using a non-EU patient cohort in terms of various performance metrics and 95% CIs: logistic regression (LR), random forest (RF), and extreme gradient boosting (XGB).

[DOCX File , 14 KB - medinform_v10i3e32949_app4.docx]

Multimedia Appendix 5

Performance of the final model derived using the EU patient cohort and validated using a non-EU patient cohort in terms of various performance metrics and 95% CIs: logistic regression (LR), random forest (RF), and extreme gradient boosting (XGB).

[DOCX File , 14 KB - medinform_v10i3e32949_app5.docx]

Multimedia Appendix 6

Distribution of deaths over time and length of intensive care unit stay.

[DOCX File , 56 KB - medinform_v10i3e32949_app6.docx]

Multimedia Appendix 7

Distribution of fraction of inspired oxygen (FiO2) for outcomes of survivors (left) and nonsurvivors (right). FiO2 was chosen as it was the variable that had the highest impact on the performance prediction, based on SHAP analysis.

[DOCX File , 66 KB - medinform_v10i3e32949_app7.docx]

Multimedia Appendix 8

Ranking of input variables of the final setup derived using the random forest–based model.

[DOCX File , 161 KB - medinform_v10i3e32949_app8.docx]

Multimedia Appendix 9

Ranking of input variables of the final setup derived using the logistic regression–based model.

[DOCX File , 126 KB - medinform_v10i3e32949_app9.docx]

Multimedia Appendix 10

List of COVIP-collaborators.

[DOCX File , 56 KB - medinform_v10i3e32949_app10.docx]

References

1. European Society of Intensive Care Medicine (ESICM), Global Sepsis Alliance (GSA), Society of Critical Care Medicine (SCCM). Reducing the global burden of sepsis: a positive legacy for the COVID-19 pandemic? Intensive Care Med 2021 Jul 16;47(7):733-736. [doi: 10.1007/s00134-021-06409-y] [Medline: 34132841]

2. Maltese G, Corsonello A, Di Rosa M, Soraci L, Vitale C, Corica F, et al. Frailty and COVID-19: a systematic scoping review. J Clin Med 2020 Jul 04;9(7):2106 [FREE Full text] [doi: 10.3390/jcm9072106] [Medline: 32635468]

3. Alkuzweny M, Raj A, Mehta S. Preparing for a COVID-19 surge: ICUs. EClinicalMedicine 2020 Aug;25:100502 [FREE Full text] [doi: 10.1016/j.eclinm.2020.100502] [Medline: 32835188]

4. Chopra V, Flanders SA, Vaughn V, Petty L, Gandhi T, McSparron JI, et al. Variation in COVID-19 characteristics, treatment and outcomes in Michigan: an observational study in 32 hospitals. BMJ Open 2021 Jul 23;11(7):e044921 [FREE Full text] [doi: 10.1136/bmjopen-2020-044921] [Medline: 34301650]

5. Mudatsir M, Fajar JK, Wulandari L, Soegiarto G, Ilmawan M, Purnamasari Y, et al. Predictors of COVID-19 severity: a systematic review and meta-analysis. F1000Res 2020 Sep 9;9:1107. [doi: 10.12688/f1000research.26186.1]

6. Flaatten H, De Lange DW, Morandi A, Andersen FH, Artigas A, Bertolini G, VIP1 study group. The impact of frailty on ICU and 30-day mortality and the level of care in very elderly patients (≥ 80 years). Intensive Care Med 2017 Dec 21;43(12):1820-1828. [doi: 10.1007/s00134-017-4940-8] [Medline: 28936626]

7. Zhao Z, Chen A, Hou W, Graham JM, Li H, Richman PS, et al. Prediction model and risk scores of ICU admission and mortality in COVID-19. PLoS One 2020 Jul 30;15(7):e0236618 [FREE Full text] [doi: 10.1371/journal.pone.0236618] [Medline: 32730358]

8. Jung C, Bruno RR, Wernly B, Wolff G, Beil M, Kelm M. Frailty as a prognostic indicator in intensive care. Dtsch Arztebl Int 2020 Oct 02;117(40):668-673. [doi: 10.3238/arztebl.2020.0668] [Medline: 33357351]

9. Wernly B, Mamandipoor B, Baldia P, Jung C, Osmani V. Machine learning predicts mortality in septic patients using only routinely available ABG variables: a multi-centre evaluation. Int J Med Inform 2021 Jan;145:104312. [doi: 10.1016/j.ijmedinf.2020.104312] [Medline: 33126059]

10. Masyuk M, Wernly B, Lichtenauer M, Franz M, Kabisch B, Muessig JM, et al. Prognostic relevance of serum lactate kinetics in critically ill patients. Intensive Care Med 2019 Jan 26;45(1):55-61. [doi: 10.1007/s00134-018-5475-3] [Medline: 30478622]

11. Bruno RR, Wernly B, Flaatten H, Fjølner J, Artigas A, Bollen Pinto B, COVIP Study Group. Lactate is associated with mortality in very old intensive care patients suffering from COVID-19: results from an international observational study of 2860 patients. Ann Intensive Care 2021 Aug 21;11(1):128 [FREE Full text] [doi: 10.1186/s13613-021-00911-8] [Medline: 34417919]

12. Leeuwenberg AM, Schuit E. Prediction models for COVID-19 clinical decision making. Lancet Digit Health 2020 Oct;2(10):e496-e497 [FREE Full text] [doi: 10.1016/S2589-7500(20)30226-0] [Medline: 32984794]

13. Perrotta F, Corbi G, Mazzeo G, Boccia M, Aronne L, D'Agnano V, et al. COVID-19 and the elderly: insights into pathogenesis and clinical decision-making. Aging Clin Exp Res 2020 Aug 16;32(8):1599-1608 [FREE Full text] [doi: 10.1007/s40520-020-01631-y] [Medline: 32557332]

14. Quer G, Arnaout R, Henne M, Arnaout R. Machine learning and the future of cardiovascular care: JACC state-of-the-art review. J Am Coll Cardiol 2021 Jan 26;77(3):300-313 [FREE Full text] [doi: 10.1016/j.jacc.2020.11.030] [Medline: 33478654]

15. Izquierdo JL, Ancochea J, Savana COVID-19 Research Group, Soriano JB. Clinical characteristics and prognostic factors for intensive care unit admission of patients with COVID-19: retrospective study using machine learning and natural language processing. J Med Internet Res 2020 Oct 28;22(10):e21801 [FREE Full text] [doi: 10.2196/21801] [Medline: 33090964]

16. Bolourani S, Brenner M, Wang P, McGinn T, Hirsch JS, Barnaby D, Northwell COVID-19 Research Consortium. A machine learning prediction model of respiratory failure within 48 hours of patient admission for COVID-19: model development and validation. J Med Internet Res 2021 Feb 10;23(2):e24246 [FREE Full text] [doi: 10.2196/24246] [Medline: 33476281]

17. Aktar S, Ahamad MM, Rashed-Al-Mahfuz M, Azad A, Uddin S, Kamal A, et al. Machine learning approach to predicting COVID-19 disease severity based on clinical blood test data: statistical analysis and model development. JMIR Med Inform 2021 Apr 13;9(4):e25884 [FREE Full text] [doi: 10.2196/25884] [Medline: 33779565]

18. Vaid A, Jaladanki SK, Xu J, Teng S, Kumar A, Lee S, et al. Federated learning of electronic health records to improve mortality prediction in hospitalized patients with COVID-19: machine learning approach. JMIR Med Inform 2021 Jan 27;9(1):e24207 [FREE Full text] [doi: 10.2196/24207] [Medline: 33400679]

19. Domínguez-Olmedo JL, Gragera-Martínez Á, Mata J, Pachón Álvarez V. Machine learning applied to clinical laboratory data in Spain for COVID-19 outcome prediction: model development and validation. J Med Internet Res 2021 Apr 14;23(4):e26211 [FREE Full text] [doi: 10.2196/26211] [Medline: 33793407]

20. Subudhi S, Verma A, Patel AB, Hardin CC, Khandekar MJ, Lee H, et al. Comparing machine learning algorithms for predicting ICU admission and mortality in COVID-19. NPJ Digit Med 2021 May 21;4(1):87. [doi: 10.1038/s41746-021-00456-x] [Medline: 34021235]

21. Syeda HB, Syed M, Sexton KW, Syed S, Begum S, Syed F, et al. Role of machine learning techniques to tackle the COVID-19 crisis: systematic review. JMIR Med Inform 2021 Jan 11;9(1):e23811 [FREE Full text] [doi: 10.2196/23811] [Medline: 33326405]

22. Pan P, Li Y, Xiao Y, Han B, Su L, Su M, et al. Prognostic assessment of COVID-19 in the intensive care unit by machine learning methods: model development and validation. J Med Internet Res 2020 Nov 11;22(11):e23128 [FREE Full text] [doi: 10.2196/23128] [Medline: 33035175]

23. Kim H, Han D, Kim J, Kim D, Ha B, Seog W, et al. An easy-to-use machine learning model to predict the prognosis of patients with COVID-19: retrospective cohort study. J Med Internet Res 2020 Nov 09;22(11):e24225 [FREE Full text] [doi: 10.2196/24225] [Medline: 33108316]

24. Burian E, Jungmann F, Kaissis G, Lohöfer FK, Spinner CD, Lahmer T, et al. Intensive care risk estimation in COVID-19 pneumonia based on clinical and imaging parameters: experiences from the Munich Cohort. J Clin Med 2020 May 18;9(5):1514 [FREE Full text] [doi: 10.3390/jcm9051514] [Medline: 32443442]

25. Shashikumar SP, Wardi G, Paul P, Carlile M, Brenner LN, Hibbert KA, et al. Development and prospective validation of a deep learning algorithm for predicting need for mechanical ventilation. Chest 2021 Jun;159(6):2264-2273 [FREE Full text] [doi: 10.1016/j.chest.2020.12.009] [Medline: 33345948]

26. Jung C, Flaatten H, Fjølner J, Bruno RR, Wernly B, Artigas A, COVIP study group. The impact of frailty on survival in elderly intensive care patients with COVID-19: the COVIP study. Crit Care 2021 Apr 19;25(1):149 [FREE Full text] [doi: 10.1186/s13054-021-03551-3] [Medline: 33874987]

27. Li Y, Sperrin M, Ashcroft DM, van Staa TP. Consistency of variety of machine learning and statistical models in predicting clinical risks of individual patients: longitudinal cohort study using cardiovascular disease as exemplar. BMJ 2020 Nov 04;371:m3919 [FREE Full text] [doi: 10.1136/bmj.m3919] [Medline: 33148619]

28. Guidet B, de Lange DW, Boumendil A, Leaver S, Watson X, Boulanger C, VIP2 study group. The contribution of frailty, cognition, activity of daily life and comorbidities on outcome in acutely admitted patients over 80 years in European ICUs:

the VIP2 study. Intensive Care Med 2020 Jan 29;46(1):57-69 [FREE Full text] [doi: 10.1007/s00134-019-05853-1] [Medline: 31784798]

29. COVIP Study. VIPSTUDY. URL: https://vipstudy.org/covip-study/ [accessed 2021-10-11]

30. Chen T, Guestrin C. XGBoost: a scalable tree boosting system. 2016 Presented at: 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining; August 2016; Machineryan Francisco, CA. [doi: 10.1145/2939672.2939785]

31. Ho TK. Random decision forests. 1995 Presented at: Third International Conference on Document Analysis and Recognition; August 14-16, 1995; Montreal p. 278-282. [doi: 10.1109/icdar.1995.598994]

32. McCullagh P, Nelder JA. Generalized Linear Models. 2nd edition. Milton Park, England: Routledge; 1989.

33. Wynants L, Van Calster B, Collins GS, Riley RD, Heinze G, Schuit E, et al. Prediction models for diagnosis and prognosis of covid-19: systematic review and critical appraisal. BMJ 2020 Apr 07;369:m1328 [FREE Full text] [doi: 10.1136/bmj.m1328] [Medline: 32265220]

34. Saito T, Rehmsmeier M. The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. PLoS One 2015 Mar 4;10(3):e0118432 [FREE Full text] [doi: 10.1371/journal.pone.0118432] [Medline: 25738806]

35. Lundberg S, Lee SI. A unified approach to interpreting model predictions. arXiv. 2017. URL: https://arxiv.org/abs/1705.07874 [accessed 2022-02-22]

36. Vink EE, Azoulay E, Caplan A, Kompanje EJO, Bakker J. Time-limited trial of intensive care treatment: an overview of current literature. Intensive Care Med 2018 Sep 22;44(9):1369-1377. [doi: 10.1007/s00134-018-5339-x] [Medline: 30136140]

37. Shrime MG, Ferket BS, Scott DJ, Lee J, Barragan-Bradford D, Pollard T, et al. Time-limited trials of intensive care for critically ill patients with cancer: how long is long enough? JAMA Oncol 2016 Jan 01;2(1):76-83 [FREE Full text] [doi: 10.1001/jamaoncol.2015.3336] [Medline: 26469222]

38. Beil M, Proft I, van Heerden D, Sviri S, van Heerden PV. Ethical considerations about artificial intelligence for prognostication in intensive care. Intensive Care Med Exp 2019 Dec 10;7(1):70 [FREE Full text] [doi: 10.1186/s40635-019-0286-6] [Medline: 31823128]

39. Jung C, Fjølner J, Bruno RR, Wernly B, Artigas A, Bollen Pinto B, COVIP Study Group. Differences in mortality in critically ill elderly patients during the second COVID-19 surge in Europe. Crit Care 2021 Sep 23;25(1):344 [FREE Full text] [doi: 10.1186/s13054-021-03739-7] [Medline: 34556171]

40. Bruno RR, Wernly B, Hornemann J, Flaatten H, FjØlner J, Artigas A, COVIP study group. Early evaluation of organ failure using MELD-XI in critically ill elderly COVID-19 patients. Clin Hemorheol Microcirc 2021;79(1):109-120. [doi: 10.3233/CH-219202] [Medline: 34487039]

41. Jung C, Bruno RR, Wernly B, Joannidis M, Oeyen S, Zafeiridis T, COVIP study group. Inhibitors of the renin-angiotensin-aldosterone system and COVID-19 in critically ill elderly patients. Eur Heart J Cardiovasc Pharmacother 2021 Jan 16;7(1):76-77 [FREE Full text] [doi: 10.1093/ehjcvp/pvaa083] [Medline: 32645153]

42. Jung C, Wernly B, Fjølner J, Bruno RR, Dudzinski D, Artigas A, the COVIP study group. Steroid use in elderly critically ill COVID-19 patients. Eur Respir J 2021 Oct 25;58(4):2100979 [FREE Full text] [doi: 10.1183/13993003.00979-2021] [Medline: 34172464]

43. Flaatten H, deLange D, Jung C, Beil M, Guidet B. The impact of end-of-life care on ICU outcome. Intensive Care Med 2021 May 19;47(5):624-625. [doi: 10.1007/s00134-021-06365-7] [Medline: 33604761]

Abbreviations

AUC: area under the receiver operating characteristic curve
CURB-65: confusion, urea, respiratory rate, blood pressure
eCRF: electronic case report form
FiO2: fraction of inspired oxygen
ICU: intensive care unit
LR: logistic regression
MIMIC-III: Medical Information Mart for Intensive Care
ML: machine learning
PPV: positive predictive value
PRC: precision-recall curve
RF: random forest
SHAP: shapely additive explanation
SOFA: Sequential Organ Failure Assessment
TLT: time-limited trials
XGB: extreme gradient boosting

XSL•FO
RenderX

XSL•FO

RenderX