

---

# JMIR Medical Informatics

---

Impact Factor (2023): 3.1  
Volume 10 (2022), Issue 2 ISSN 2291-9694 Editor in Chief: Christian Lovis, MD, MPH, FACMI

---

## Contents

### Reviews

- State-of-the-art Dashboards on Clinical Indicator Data to Support Reflection on Practice: Scoping Review (e32695)  
Bernard Bucalon, Tim Shaw, Kerri Brown, Judy Kay. . . . . 3
- The Science of Learning Health Systems: Scoping Review of Empirical Research (e34907)  
Louise Ellis, Mitchell Sarkies, Kate Churruca, Genevieve Dammary, Isabelle Meulenbroeks, Carolynn Smith, Chiara Pomare, Zeyad Mahmoud, Yvonne Zurynski, Jeffrey Braithwaite. . . . . 19

### Viewpoints

- Operationalizing and Implementing Pretrained, Large Artificial Intelligence Linguistic Models in the US Health Care System: Outlook of Generative Pretrained Transformer 3 (GPT-3) as a Service Model (e32875)  
Emre Sezgin, Joseph Sirrianni, Simon Linwood. . . . . 36
- A Free, Open-Source, Offline Digital Health System for Refugee Care (e33848)  
Henry Ashworth, Senan Ebrahim, Hassaan Ebrahim, Zahra Bhaiwala, Michael Chilazi. . . . . 43

### Original Papers

- Evaluation of Natural Language Processing for the Identification of Crohn Disease–Related Variables in Spanish Electronic Health Records: A Validation Study for the PREMONITION-CD Project (e30345)  
Carmen Montoto, Javier Gisbert, Iván Guerra, Rocío Plaza, Ramón Pajares Villarroja, Luis Moreno Almazán, María López Martín, Mercedes Domínguez Antonaya, Isabel Vera Mendoza, Jesús Aparicio, Vicente Martínez, Ignacio Tagarro, Alonso Fernández-Nistal, Lea Canales, Sebastian Menke, Fernando Gomollón, PREMONITION-CD Study Group. . . . . 49
- Traditional Machine Learning Models and Bidirectional Encoder Representations From Transformer (BERT)–Based Automatic Classification of Tweets About Eating Disorders: Algorithm Development and Validation Study (e34492)  
José Benítez-Andrades, José-Manuel Alija-Pérez, Maria-Esther Vidal, Rafael Pastor-Vargas, María García-Ordás. . . . . 58
- Identification of Prediabetes Discussions in Unstructured Clinical Documentation: Validation of a Natural Language Processing Algorithm (e29803)  
Jessica Schwartz, Eva Tseng, Nisa Maruthur, Masoud Rouhizadeh. . . . . 71

Reduction of Platelet Outdating and Shortage by Forecasting Demand With Statistical Learning and Deep Neural Networks: Modeling Study ( <a href="#">e29978</a> ) Maximilian Schilling, Lennart Rickmann, Gabriele Hutschenreuter, Cord Spreckelsen. ....	80
Comparison of the Erectile Dysfunction Drugs Sildenafil and Tadalafil Using Patient Medication Reviews: Topic Modeling Study ( <a href="#">e32689</a> ) Maryanne Kim, Youran Noh, Akihiko Yamada, Song Hong. ....	94
Electricity, Computing Hardware, and Internet Infrastructures in Health Facilities in Sierra Leone: Field Mapping Study ( <a href="#">e30040</a> ) Emeka Chukwu, Lalit Garg, Edward Foday, Abdul Konomanyi, Royston Wright, Francis Smart. ....	106
Disease Progression of Hypertrophic Cardiomyopathy: Modeling Using Machine Learning ( <a href="#">e30483</a> ) Matej Pulin, Tim Smole, Bojan Žunkovi, Enja Kokalj, Marko Robnik-Šikonja, Matjaž Kukar, Dimitrios Fotiadis, Vasileios Pezoulas, Nikolaos Tachos, Fausto Barlocco, Francesco Mazzarotto, Dejana Popovi, Lars Maier, Lazar Velicki, Iacopo Olivotto, Guy MacGowan, Djordje Jakovljevi, Nenad Filipovi, Zoran Bosni. ....	115
Early Identification of Maternal Cardiovascular Risk Through Sourcing and Preparing Electronic Health Record Data: Machine Learning Study ( <a href="#">e34932</a> ) Nawar Shara, Kelley Anderson, Noor Falah, Maryam Ahmad, Darya Tavazoei, Justin Hughes, Bethany Talmadge, Samantha Crovatt, Ramon Dempers. ....	133
Development of Deep Learning Models for Predicting In-Hospital Mortality Using an Administrative Claims Database: Retrospective Cohort Study ( <a href="#">e27936</a> ) Hiroki Matsui, Hayato Yamana, Kiyohide Fushimi, Hideo Yasunaga. ....	142
Panic Attack Prediction Using Wearable Devices and Machine Learning: Development and Cohort Study ( <a href="#">e33063</a> ) Chan-Hen Tsai, Pei-Chen Chen, Ding-Shan Liu, Ying-Ying Kuo, Tsung-Ting Hsieh, Dai-Lun Chiang, Feipei Lai, Chia-Tung Wu. ....	153
Early Prediction of Functional Outcomes After Acute Ischemic Stroke Using Unstructured Clinical Text: Retrospective Cohort Study ( <a href="#">e29806</a> ) Sheng-Feng Sung, Cheng-Yang Hsieh, Ya-Han Hu. ....	166
The Application and Comparison of Machine Learning Models for the Prediction of Breast Cancer Prognosis: Retrospective Cohort Study ( <a href="#">e33440</a> ) Jialong Xiao, Miao Mo, Zezhou Wang, Changming Zhou, Jie Shen, Jing Yuan, Yulian He, Ying Zheng. ....	179
Automatically Explaining Machine Learning Predictions on Severe Chronic Obstructive Pulmonary Disease Exacerbations: Retrospective Cohort Study ( <a href="#">e33043</a> ) Siyang Zeng, Mehrdad Arjomandi, Gang Luo. ....	190
Exploring Quality Differences in Telemedicine Between Hospital Outpatient Departments and Community Clinics: Cross-sectional Study ( <a href="#">e32373</a> ) Noora Alhajri, Mecit Simsekler, Buthaina Alfalasi, Mohamed Alhashmi, Hamda Memon, Emma Housser, Abdulhamid Abdi, Nahed Balalaa, Maryam Al Ali, Raghda Almaashari, Shammah Al Memari, Farida Al Hosani, Yousif Al Zaabi, Shereena Almazrouei, Hamed Alhashemi. ....	213

Review

# State-of-the-art Dashboards on Clinical Indicator Data to Support Reflection on Practice: Scoping Review

Bernard Bucalon<sup>1,2\*</sup>, BDesComp; Tim Shaw<sup>2,3\*</sup>, BSc, PhD; Kerri Brown<sup>2,4\*</sup>, BA, MA, PhD; Judy Kay<sup>1,2\*</sup>, BSc, PhD

<sup>1</sup>Human Centred Technology Cluster, School of Computer Science, The University of Sydney, Darlington, Australia

<sup>2</sup>Practice Analytics, Digital Health Cooperative Research Centre, Sydney, Australia

<sup>3</sup>Research in Implementation Science and e-Health Group, Faculty of Medicine and Health, The University of Sydney, Sydney, Australia

<sup>4</sup>Professional Practice Directorate, The Royal Australasian College of Physicians, Sydney, Australia

\*all authors contributed equally

**Corresponding Author:**

Bernard Bucalon, BDesComp  
Human Centred Technology Cluster  
School of Computer Science  
The University of Sydney  
Level 3, Computer Science Building J12  
1 Cleveland St  
Darlington, NSW 2008  
Australia  
Phone: 61 2 8627 0010  
Email: [bernard.bucalon@sydney.edu.au](mailto:bernard.bucalon@sydney.edu.au)

## Abstract

**Background:** There is an increasing interest in using routinely collected eHealth data to support reflective practice and long-term professional learning. Studies have evaluated the impact of dashboards on clinician decision-making, task completion time, user satisfaction, and adherence to clinical guidelines.

**Objective:** This scoping review aims to summarize the literature on dashboards based on patient administrative, medical, and surgical data for clinicians to support reflective practice.

**Methods:** A scoping review was conducted using the Arksey and O'Malley framework. A search was conducted in 5 electronic databases (MEDLINE, Embase, Scopus, ACM Digital Library, and Web of Science) to identify studies that met the inclusion criteria. Study selection and characterization were performed by 2 independent reviewers (BB and CP). One reviewer extracted the data that were analyzed descriptively to map the available evidence.

**Results:** A total of 18 dashboards from 8 countries were assessed. Purposes for the dashboards were designed for performance improvement (10/18, 56%), to support quality and safety initiatives (6/18, 33%), and management and operations (4/18, 22%). Data visualizations were primarily designed for team use (12/18, 67%) rather than individual clinicians (4/18, 22%). Evaluation methods varied among asking the clinicians directly (11/18, 61%), observing user behavior through clinical indicators and use log data (14/18, 78%), and usability testing (4/18, 22%). The studies reported high scores on standard usability questionnaires, favorable surveys, and interview feedback. Improvements to underlying clinical indicators were observed in 78% (7/9) of the studies, whereas 22% (2/9) of the studies reported no significant changes in performance.

**Conclusions:** This scoping review maps the current literature landscape on dashboards based on routinely collected clinical indicator data. Although there were common data visualization techniques and clinical indicators used across studies, there was diversity in the design of the dashboards and their evaluation. There was a lack of detail regarding the design processes documented for reproducibility. We identified a lack of interface features to support clinicians in making sense of and reflecting on their personal performance data.

(*JMIR Med Inform* 2022;10(2):e32695) doi:[10.2196/32695](https://doi.org/10.2196/32695)

**KEYWORDS**

practice analytics dashboards; data visualization; reflective practice; professional learning; mobile phone

## Introduction

### Background

Dashboards have evolved from single-view reporting information based on large raw data sets to customizable interfaces with multiple views and purposes, such as communication, learning, motivation, monitoring, and decision support [1]. The use of dashboards in many clinical settings has been well established [2]. Studies on these dashboards have focused on patient monitoring and clinical decision support using electronic medical records (EMRs), electronic audit and feedback (e-A&F) systems based on quality and safety standards, and management dashboards to support the day-to-day operations of departments. Evaluations of clinical dashboards tend to evaluate accuracy (decision-making), efficiency (time-to-task completion), usability (user satisfaction), and adherence to guidelines (quality and safety) [3-5].

There are known organizational, cultural, and technical issues with collecting and reporting on clinical indicators [6], sometimes referred to as quality or performance indicators. Despite the limitations of clinical indicator data, health professionals' attitudes suggest that there is an appetite for easier and timely access to routinely collected clinical indicator data for personalized performance feedback [7]. Mainz et al [8] categorizes clinical indicators as *structural*, *process*, or *outcome indicators*. Structural indicators describe the type and number of resources by a health system or organization to deliver care, for example, the number of staff, patients, beds, and supplies. Process indicators measure the activities and tasks in patient care episodes, for example, patients were assessed by a physician within 24 hours of referral, and patients were treated according to clinical guidelines. Outcome indicators are states of health or events that follow care, which may be affected by health care. Mainz et al [8] proposes that outcome indicators are usually related to death, disease, discomfort, disability, and dissatisfaction. Clinical indicators can also be categorized as *generic* or *disease-specific*. Generic indicators measure aspects of care that are relevant to most patients (length of stay, readmissions, and late discharges). Disease-specific indicators are diagnosis-specific and measure specific aspects of diseases and conditions (hip fractures after the second operation and patients with lung cancer who are alive 30 days after surgery).

Reflective practice and lifelong learning are central to continuing professional development (CPD) frameworks mandated by medical boards around the world [9]. Participation in CPD programs ensures that medical specialists meet the standards required to maintain their specialist registration. Although board examinations and work-based assessments certify the initial competence of medical graduates, practicing clinicians require ongoing self-assessment to maintain standards and identify improvement needs [10]. Professional development frameworks often include references to the use of practice data for clinical audits and reflection [11,12].

Dashboards are commonly used for audit and feedback (A&F), an established process for improving professional practice by reviewing data based on existing benchmarks in the quality and safety literature [13]. Although there have been some successes,

existing studies on e-A&F dashboards show that evidence is limited in terms of effectiveness for improving performance [14]. Furthermore, studies on dashboards designed to support clinician reflective practice and lifelong professional learning are scarce and heterogeneous [15].

This work aims to fill the gap in the literature on the use of data from disparate clinical sources to generate new insights that lead to practice reflection by clinicians.

e-A&F dashboards address known questions about clinical performance, whereas clinical *practice reflection* dashboards focus on presenting routinely collected data to clinicians to engage with and reflect on and to reveal new questions about their individual and wider team practice.

There has been some emerging literature on clinical practice reflection dashboards designed to support the reflective practice of clinicians [16]. This scoping review will summarize the literature on dashboards that support the reflective practice of clinicians and systematically map the features and outcomes of published interventions.

### Objectives

This scoping review aims to systematically map the different characteristics of feedback interfaces that support clinicians in reflecting on their practice. The data extracted from the included studies will provide insight into why the interfaces were created, how they were designed and evaluated, and what were the reported outcomes.

The scoping review was guided by the following 6 research questions (RQs):

- RQ1: What was the purpose of the performance feedback interfaces?
- RQ2: What clinical indicators were used and how are they visualized?
- RQ3: How were the interfaces designed?
- RQ4: What were the methods used to evaluate the interfaces?
- RQ5: How successful have the interfaces been?
- RQ6: What are the key design considerations for developing future interfaces?

## Methods

### Overview

The scoping review process was conducted following the methodology and guidelines by Arksey and O'Malley [17]. The process is outlined in 6 steps as follows: identifying the RQ; identifying the relevant studies; study selection; charting the data; collating, summarizing, and reporting results; and consultation.

To ensure the quality of the studies, the review only included studies published in peer-reviewed journals that had the full text available. Additional quality analysis was not conducted on the included studies, as quality assessment is not a requirement for a scoping review [18], and there are no established criteria to evaluate the quality of clinical dashboard

studies. Quality assessment was not conducted to ensure that lessons were gained from a diverse range of work.

### Search Strategy

The search strategy was developed in consultation with the university librarian, using the *Population-Concept-Context* mnemonic [19].

The target population included any medical specialist as defined by the Australian Health Practitioner Regulation Agency registry of medical specialties and subspecialties, general practitioners (primary care physicians), and registrars (residents) in specialist medical training programs.

The review explored the concept of the use of clinical indicators to provide insight into a clinician's own practice. Synonyms were generated for the search term clinical indicator, such as *quality indicator* and *performance indicator*. Generic terms for possible data sources for clinical indicators included search terms, such as *administrative*, *medical*, and *surgical data*. An additional concept focused on the intervention used in the study, that is, the feedback user interface. Synonyms for search terms included the following: *dashboard*, *visualization*, and *report*. Search terms, such as *feedback* and *reflection* were specifically not included to maximize the breadth of the search. We ensured that performance feedback and reflection dashboard studies were still captured in the search using the *clinical indicator* and *interface* search terms.

Peters et al [19] defines a context in terms of geographic location, setting, or cultural factors. No specific search terms were used for context, as there were no requirements related to the country of study, and hospital setting, such as public, private, inpatient, outpatient, rural, remote, or metropolitan.

The search strategy was developed by BB and was refined based on feedback from all the authors and the university librarian (JG). The search strategy was translated into a search query (Multimedia Appendix 1) and conducted on the following electronic databases: MEDLINE, Embase, Scopus (which includes IEEE Xplore and ScienceDirect), ACM Digital Library, and Web of Science. The electronic databases were selected to ensure coverage of clinical dashboard studies published in the fields of health informatics, data visualizations, and human-computer interaction research.

### Search and Study Selection

The initial search was conducted by BB. Articles were retrieved based on an agreed-upon search strategy. Next, BB screened all the retrieved article titles and abstracts against the inclusion and exclusion criteria. Concurrently, CP screened a random selection of 4.99% (184/3685) of the retrieved articles. BB and CP discussed the conflicts generated during the screening. The authors agreed to proceed to a full-text review after the first review returned 4.3% (8/184) conflicts from the abstracts screened.

Studies were eligible for this review if they met the following criteria: the study provided a medical practitioner with access to clinical indicator data to receive feedback on their performance, included details on the design and implementation of the interface, included information on the interface features

(visual and functions), described the evaluation methods used, and was published in English in a peer-reviewed journal between 2010 and 2020.

Articles were excluded if the study participants were in medical school, as medical students were not considered as professional learners. Articles were also excluded if the intervention was designed for public health physicians and researchers, as the interface was concerned with data about communities and populations. If the full text was not available (eg, conference abstracts), the article was excluded. Articles that grouped a variety of medical practitioners were included. However, they were excluded if it was unclear which reported data and findings related to participants in our inclusion criteria.

### Data Extraction

Data were extracted from articles retrieved by BB and then reviewed by CP to mitigate bias. Table 1 maps the RQs to the descriptive data extracted from the included studies.

For RQ1, the purpose of each interface was extracted because there were no specialty or subspecialty restrictions on the search. By identifying the stated purpose and aims of each dashboard, we could better compare similar implementations. For example, dashboards for managing day-to-day operations are compared with dashboards for clinical quality improvement.

To understand the data presented in each interface, RQ2 extracted the names of clinical indicators (eg, length of stay), the data source (eg, EMR and clinical registry), and the technology used by the platforms. RQ2 also captured how the data were presented and the features of the interface by extracting design details, such as the data visualization types (eg, bar chart), interactivity (eg, zooming and filtering), and individual versus team views. We also identified the intended frequency of use of the interface. Kay et al [20] describes 2 mental systems that work differently and drive the way people think. System 1 performs *fast* intuitive thinking, which is automatic but can lead to bias and errors. In contrast, system 2 performs *slow* rational and logical thinking that is conscious and can override the initial insights acquired by system 1 [20]. We define *fast* use as at a glance or daily use. We define *slow* use as longer than a day (eg, weekly or monthly use).

RQ3 extracted the design process used in each study, as there is value in understanding how the interfaces were designed. We anticipate that the design approaches and methods used in these studies could be helpful for future researchers to design similar interfaces.

To address RQ4, we expected to see a diverse range of research methods conducted across controlled laboratory and authentic hospital settings. RQ5 then looked to assess which interfaces were effective in terms of usability, changes in practice, and patient outcomes. RQ4 and RQ5 together allow us to gauge the success of the studies in achieving their stated goals.

Finally, RQ6 identified the key factors to consider when designing interfaces to support the reflective practice of clinicians. RQ6 summarizes the practice points and recommendations proposed by the included studies.

**Table 1.** Research questions (RQs) and data planned to be extracted from included studies.

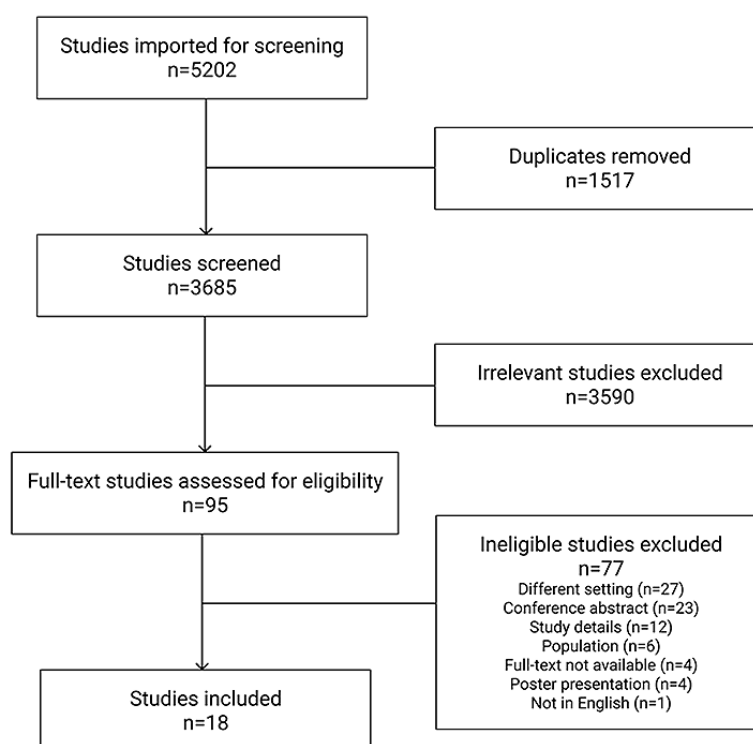
RQ		Data extracted
RQ1	What was the purpose of the performance feedback interfaces?	<ul style="list-style-type: none"> <li>Stated purpose and aims</li> </ul>
RQ2	What clinical indicators were used and how are they visualized?	<ul style="list-style-type: none"> <li>Clinical indicators</li> <li>Visualization elements</li> <li>Frequency of intended use</li> <li>Individual or team use</li> <li>Static or interactive features</li> <li>Data source</li> <li>Technology</li> </ul>
RQ3	How were the interfaces designed?	<ul style="list-style-type: none"> <li>Design process</li> </ul>
RQ4	What are the methods used to evaluate the interfaces?	<ul style="list-style-type: none"> <li>Evaluation methods</li> <li>Laboratory vs in-the-wild settings</li> </ul>
RQ5	How successful have the interfaces been?	<ul style="list-style-type: none"> <li>Reported results and outcomes</li> <li>Strengths and limitations</li> </ul>
RQ6	What are the key design considerations for developing future interfaces?	<ul style="list-style-type: none"> <li>Practice points</li> <li>Recommendations</li> </ul>

## Results

### General Characteristics of the Included Studies

The following section summarizes the general characteristics of the included studies, such as the year of publication, location of publication, citation trends, country of origin, specialty of participants, and study duration ([Multimedia Appendix 2 \[21-38\]](#)).

[Figure 1](#) shows the flow of articles from the identification, screening, and final inclusion. The original search conducted in November 2020 yielded 3685 potentially relevant citations after duplicates ( $n=1517$ ) were removed. After title and abstract screening, 2.58% (95/3685) citations met the eligibility criteria, and the corresponding full-text articles were procured for full-text review. After reviewing all the full-text articles, 81% (77/95) studies were excluded according to the inclusion and exclusion criteria; 19% (18/95) dashboard studies remained and were included in the analysis.

**Figure 1.** Flow diagram of search and selection studies.

Of the 77 excluded studies, 27 (35%) were excluded for having a different setting. Studies were identified as having a different setting if the interface was not used to provide feedback on individual or team performance, for example, used outside a hospital or clinic environment, such as public health researchers. In all, 8% (6/77) of the studies were excluded as the target participants of the study did not meet the inclusion criteria, for example, nurses, pharmacists, and medical researchers. Overall, 16% (12/77) of the studies were excluded owing to insufficient study details to be extracted for analysis. Studies were also omitted if the full study text was not accessible, for example, conference abstracts (23/77, 30%) and poster presentations (4/77, 5%). Of the 77 excluded studies, 4 (5%) could not be retrieved, and 1 (1%) study was excluded because it was not published in English.

All studies were published between 2010 and 2020, with 72% (13/18) published after 2015. Most of the citations (181/208, 87%) occurred after 2015. Linder et al [21] contributed 27.9% (58/208) of all the citations in the studies between 2010 and 2020.

Of the 18 selected studies, 3 (17%) studies [22-24] cited another study on an electronic health record (EHR) dashboard to improve antibiotic prescription [21]. Laurent et al [23] also cited a maternity dashboard pilot study [25]. Schall et al [24] cited an earlier study by the same authors on the evaluation of a health care information technology (HIT) dashboard based on quality indicators [26].

The countries of origin of these studies are summarized in Table 2. Most of the studies were conducted in English-speaking countries, with 55% (10/18) originating from the United States.

**Table 2.** Country of origin from included studies (N=18).

Country of origin	Count, n (%)	References
United States	10 (56)	[21,22,24,26-32]
Australia	2 (11)	[33,34]
Other (Canada, France, the Netherlands, Oman, Sweden, and United Kingdom)	6 (33)	[23,25,35-38]

By looking at the study participants, we could identify the specialty and subspecialty groups that use interfaces to engage with data about their performance. The participants of the included studies came from 11 distinct medical specialties or subspecialties, with 22% (4/18) of dashboard studies focusing on primary care physicians (or general practitioners). Studies have also evaluated dashboards for anesthesia (3/18, 17%). Of the 18 included studies, 3 (17%) studies did not specify a particular specialty or subspecialty of the participants, and 1 (6%) study included registrars (residents) who were still in specialist training programs.

In all, 50% (9/18) of the studies did not specify the duration of the evaluation. Of these studies, 56% (5/9) of the studies were deployed and evaluated in real-world hospital environments, and 44% (4/9) of the studies were conducted in controlled laboratory settings. The evaluation study duration ranged between 2 months [23] and 42 months [27].

**Table 3.** Purpose of included dashboard studies grouped by category (N=18).<sup>a</sup>

Purpose	Count, n (%)	References
Performance improvement	9 (50)	[27-31,33,36-38]
Quality and safety	6 (33)	[21,23-25,32,35]
Management and operations	4 (22)	[22,23,26,34]

<sup>a</sup>Included studies may be in more than 1 category.

## RQ2: Common Features

The following section summarizes the common clinical indicators used across the dashboards, how the indicators were presented to the end users, where the indicators were sourced, and the technology platform details. By identifying the lower-level data used and the functionality of each dashboard,

## RQ1: Purpose

As the studies were conducted across a range of medical specialties, the purpose for each dashboard was also diverse.

Table 3 shows that the clinical dashboards in the review fell evenly across 3 categories. Performance improvement dashboards aim to present data to an individual or team to reflect on their practice and identify areas to change. Quality and safety dashboards track the agreed-upon clinical guidelines and benchmarks. They can be modeled with existing clinical practice improvement models, such as Plan-Do-Study-Act [39]. Management and operations dashboards are targeted to administrators and directors of departments to support the day-to-day functions of health care services. Laurent et al [23] was categorized as supporting quality and safety as well as management and operations.

we can see how the researchers aimed to fulfill the purpose of their feedback interface.

## Clinical Indicators

As shown in Table 4, the use of clinical indicators varied across the studies. The study by Clark et al [34] was the only study that evaluated a dashboard that presented structural indicators to clinicians, such as consultant workload and bed availability.

Process indicators were used in all but one of the studies. The most used generic indicators across the studies included the following: length of stay (7/18, 39%), readmission (4/18, 22%), and discharge (3/18, 17%), whereas % acute respiratory infections (ARI) visits with antibiotics, Lymphedema index

(L-Dex), and Number of Atrial fibrillation (AF) diagnosis over time were examples of specialty-specific indicators.

Of the 18 studies, 3 (17%) studies presented outcome indicators—mortality [23], patient complaints [25], and patient satisfaction [30].

**Table 4.** Clinical indicators by type from included studies (N=18).<sup>a</sup>

Clinical indicators	Count, n (%)	References
<b>Classification</b>		
Structural	1 (6)	[34]
Process	17 (94)	[21-38]
Outcome	5 (28)	[23,26,30,36,38]
<b>Specificity</b>		
Generic	15 (83)	[22-32,34-37]
Disease-specific	3 (17)	[21,33,38]

<sup>a</sup>Included studies may have more than 1 type of clinical indicator.

### Dashboard Presentation

The types of visualization used to present the underlying clinical indicators are summarized in Table 5. A combination of bar charts, tables, and line charts were used in 50% (9/18) of the studies. Gude et al [36] and Weiner et al [22] used all the 3 techniques.

Table 6 shows that most (12/18, 67%) of the dashboards evaluated were intended for team use, whereas 22% (4/18) of the dashboards were for individual use. Of the 18 dashboards evaluated, 2 (11%) dashboards were designed for both team and individual use. Clinicians work in specialty care teams,

multidisciplinary teams, and as individual consultants; therefore, dashboard interfaces should show the relevant data depending on the setting.

The intended frequency of use of the dashboards was evenly split between fast and slow use as shown in Table 7. One urology dashboard was designed specifically for rapid or at-a-glance use [37]. In all, 17% (3/18) of the dashboards were designed for slow use and were reviewed every month [25,27,28]. Overall, 22% (4/18) of the included studies designed dashboards for slow use but did not specify the exact cadence for reviewing data [21,33,36,38].

**Table 5.** Dashboard visualization elements used in included studies (N=18).<sup>a</sup>

Visualization elements	Count, n (%)	References
Bar chart including histogram	10 (56)	[21-23,27,29,32-34,36,37]
Table	9 (50)	[22-27,31,36,38]
Line chart	9 (50)	[22,29,31,32,34-38]
Scatter plot	1 (6)	[35]
Meter	1 (6)	[22]
Radar including radial or spider-web	1 (6)	[30]
Pie chart including donuts or rings	1 (6)	[29]

<sup>a</sup>Included studies may have more than 1 visualization element.

**Table 6.** Dashboard designed for team or individual use (N=18).

Use	Count, n (%)	References
Team	12 (67)	[22-29,32,34,37,38]
Individual	4 (22)	[21,30,35,36]
Both	2 (11)	[31,33]



**Table 7.** Dashboard studies designed for fast or slow use (N=18).

Use	Count, n (%)	References
<b>Fast</b>		
Daily	8 (44)	[22-24,26,29,31,32,34]
Rapid or at-a-glance	1 (6)	[37]
<b>Slow</b>		
Weekly	1 (6)	[30]
Monthly	3 (17)	[25,27,28]
Quarterly	1 (6)	[35]
No details	4 (22)	[21,33,36,38]

### Data Sources

Of the 18 included studies, 5 (28%) studies were conducted on top of the EMR and EHR systems, 4 (22%) studies were integrated with an existing data warehouse within the hospital infrastructure, 2 (11%) studies were integrated with the patient administration systems within the hospital, and 1 (6%) study used data from a clinical registry. In all, 44% (8/18) of the studies did not specify the data source used to implement the dashboard solution.

### Technology

Overall, 22% (4/18) of the dashboard studies leveraged web development technologies, such as HTML, Cascading Style Sheet, and JavaScript. Open-source libraries, such as jQuery (OpenJS Foundation), D3.js (Mike Bostock), and HighCharts (Highsoft AS) were also used. Of the 18 studies, 4 (22%) studies used out-of-the-shelf enterprise solutions (SAS, Tableau, and

Qlikview). In total, 11% (2/18) of the dashboard studies presented data using Microsoft Excel. In all, 33% (6/18) of the included studies did not specify the technology tools and platforms used to implement the dashboard solution.

Dashboards can be interactive, allowing users to engage with the data in multiple ways rather than a single static view. Interactive dashboards enable users to drilldown to gain background information, show comparisons, and highlight anomalies in the data visualizations [40]. Shneiderman et al [41] has described common features of advanced graphical user interfaces, including an overview of the entire collection of data, zooming into interesting items, filtering out uninteresting items, and retrieving additional details on demand.

Table 8 shows that most (15/18, 83%) of the dashboards were interactive, 11% (2/18) of the dashboards were static, and 6% (1/18) of the studies did not provide details on whether the dashboard was interactive or static [28].

**Table 8.** Dashboard studies designed to be interactive or static (N=18).

Interface design	Count	References
Interactive	15 (83)	[21-24,26,27,29,31-38]
Static	2 (11)	[25,30]
No details	1 (6)	[28]

### RQ3: Design Process

Only 56% (10/18) of the studies provided details on the design process used. User-centered design (3/18, 17%), co-design (2/18, 11%), and iterative processes (2/18, 11%) were the specific approaches mentioned in the papers. Mulhall et al [35] used both co-design and user-centered design approaches. An iterative and user-centered design was used by Stattin et al [37]. Although the remaining 44% (8/18) of the studies had no details of the design processes, 55% (10/18) of the studies did report details used a diverse range of methods, including focus groups, interviews, workshops, and process mapping.

### RQ4: Evaluation Methods

The evaluation methods, grouped by type, are listed in Table 9. A mix of quantitative and qualitative research methods were

used. Majority of the studies (10/18, 57%) quantitatively evaluated the impact of the dashboards. Data were primarily sourced from EMRs, clinical registries, and patient administrative systems. Questionnaires, such as pre- and postsurveys, standardized single ease questions (SEQ) and system usability scale (SUS) were used in 44% (8/18) of the studies. Methods used in the remaining studies included analysis of access logs (3/18, 17%), formal cluster randomized control trials (2/18, 11%), case studies (2/18, 11%), interviews, the think-aloud protocol, heuristic evaluation, and eye tracking.

In terms of the evaluation setting, most of the studies (13/18, 72%) were conducted in authentic settings, such as in the emergency department, primary care clinics, or hospital inpatient wards, whereas the remaining studies (4/18, 22%) were conducted in controlled laboratory settings.

**Table 9.** Evaluation methods used by included studies (N=18).<sup>a</sup>

Method	Count, n (%)	References
<b>Asking the users</b>		
Questionnaires or surveys	9 (50)	[23,24,26,28,31,33,35,36,38]
Interviews	2 (11)	[33,38]
<b>Evaluating user behavior</b>		
eHealth data analysis	10 (56)	[21,22,27-31,34,36,37]
System usage log analysis	4 (22)	[21,27,29,35]
<b>Evaluating usefulness of the interface</b>		
Expert method	1 (6)	[26]
Usability user study	3 (17)	[24,33,38]

<sup>a</sup>Included studies may have more than 1 evaluation method.

### RQ5: Reported Outcomes

The reported outcomes of each of the included dashboard studies are summarized further ([Multimedia Appendix 3 \[21-38\]](#)). The methods were grouped by (1) direct feedback from end users, (2) data analysis of the underlying eHealth data, (3) data analysis of platform use logs, (4) expert usability evaluation, and (5) usability testing with end users.

#### Asking the Users: Questionnaire, Survey, and Interview

Studies that asked for feedback on the dashboard directly from end users using standard questionnaires, surveys, and interviews are summarized in [Table 10](#).

Of the 18 included studies, 5 (28%) studies used a standardized questionnaire to gauge the individuals' assessment of dashboard usability. SUS is a validated questionnaire that measures users'

overall satisfaction with a graphical user interface [42]. The questionnaire is interface agnostic and consists of 10 items with total scores ranging from 0 to 100 [43]. In all, 17% (3/18) of the included studies reported high mean SUS scores of 82.6 (SD 11.5) [23], 83 (SD 7.6) [26], and 87.5 (SD 9.6) [24]. Overall, 6% (1/18) of the studies reported a median SUS score of 73.0 (SD 15.0) [38].

In addition to the SUS questionnaire, Schall et al [24] also conducted a Post-Study System Usability Questionnaire (PSSUQ). The PSSUQ consists of 19 items that measure users' perceived satisfaction with a product [44]. The questionnaire consists of three subscales as follows: system usefulness, information quality, and interface quality. The study found an overall mean PSSUQ score of 1.7 (SD 0.5) with subscale scores of 1.5 (SD 0.4), 1.8 (SD 0.8), and 1.8 (SD 0.8)—suggesting the dashboard had good usability.

**Table 10.** Reported results from standardized questionnaires, surveys, and interviews (N=18).

Evaluation method	Reported outcomes	References
Standardized questionnaire	<ul style="list-style-type: none"> <li>Mean SUS<sup>a</sup> score of at least 73.0 across 5 studies (range 73.0-87.5).</li> <li>PSSUQ<sup>b</sup> score of 1.7 (SD 0.5).</li> <li>All tasks rated median SEQ<sup>c</sup> score of 1 (very easy) or 2 (easy).</li> </ul>	[23,24,26,33,38]
Survey	<ul style="list-style-type: none"> <li>Respondents had favorable responses to the dashboards (range 72-79).</li> <li>Respondents stated the data were actionable (range 48-69).</li> <li>Respondents felt the data improve their practice (range 64-98).</li> </ul>	[28,29,31,35]
Interview	<ul style="list-style-type: none"> <li>Interviewees were interested and enthusiastic about the individual patient dashboard.</li> <li>Interviewees were generally excited to have the opportunity to see the cohort dashboard but commented on its complexity.</li> <li>Interviewees were generally positive about the clinical performance summary, patient lists, suggested actions, and detailed patient-level information views.</li> <li>Interviewees identified improvements on the clinical performance summaries view (eg, inclusion of CIs with differing guidance was confusing).</li> </ul>	[33,38]

<sup>a</sup>SUS: system usability scale.

<sup>b</sup>PSSUQ: Post-Study System Usability Questionnaire.

<sup>c</sup>SEQ: single ease question.

Janssen et al [33] asked 5 participants to rate the ease of completing tasks on the lymphedema dashboards using a 7-point SEQ [45]. All tasks on the individual dashboard received a median SEQ rating of either 1 (very easy) or 2 (easy). Similarly,

all tasks on the cohort dashboard received a median SEQ rating of 1 or 2. The last task on the cohort dashboard was attempted by only 3 participants and received a median SEQ rating of 3.

User feedback surveys were conducted in 17% (3/18) of the studies. Mulhall et al [35] surveyed 316 family physicians who used the quality improvement dashboard in long-term care practice. The overall quality of the dashboard was rated as *good* (45%) and *very good* (34%), and 69% of physicians said they were *likely* or *very likely* to implement one of the suggested changes.

Khanna et al [28] surveyed 48 primary and specialty care practices on their perceptions of a practice transformation analytics dashboard as a tool to present data that are actionable in health care design. The study found that 96% of surveyed practices reported having previously reviewed their cost data, 72% had favorable responses to the dashboard, and 48% found dashboard data actionable (n=25).

Ehrenfeld et al [31] evaluated the perceptions of anesthesia residents on a performance feedback dashboard. The study found that 91% of respondents said they would like to receive a systematic review of practice performance data every 1 to 4 weeks (n=48), whereas 98% of resident respondents said they could improve in at least one and often multiple areas. Only 10% of the respondents believed that they were compliant in all 6 areas listed. All respondents, except 1, noted that they would like to receive feedback in some electronic form, for example, emails, websites, and smartphones.

### ***Evaluating User Behavior Through eHealth Data Analysis***

Table 11 summarizes the key results related to changes in the underlying clinical indicators and the dashboard use logs.

**Table 11.** Reported outcomes from data analysis of eHealth data and system use logs.

Evaluation method	Reported outcomes	References
eHealth data analysis	<ul style="list-style-type: none"> <li>2 out of 9 studies evaluating eHealth data reported positive changes to CI data.</li> <li>2 out of 9 studies reported no change to CI data.</li> </ul>	[21,22,27,29,30,34-37]
System use log data analysis	<ul style="list-style-type: none"> <li>&gt;50% of participants viewed the dashboard in 2 studies (range 28-50).</li> <li>A median of 55 views from 30 users was observed in 1 study.</li> </ul>	[21,29,35]

Hester et al [27] observed improvements in emergency department balancing measures, which included a higher emergency department discharge rate (70.7% vs 72.8%;  $P=.05$ ), lower charges (ratio 1:0.86;  $P<.001$ ), shorter length of stay (2.9 hours vs 2.6 hours;  $P=.001$ ), and lower 7-day revisit rates (15.4% vs 11.6%;  $P<.001$ ). Inpatient charges decreased (ratio 1:1.14;  $P=.01$ ), but the length of stay and readmission remained stable.

Patel et al [29] observed that the composite discharge mix index improved during the 5-month study period; they observed a 79.3% completion rate in the intervention group (n=537) compared with 63.2% in the control group (n=516).

In 53.8% of the cases (n=288), Gude et al [36] observed that intensive care specialists overestimated their clinical performance, whereas in 13.5% of the cases, they underestimated their performance. Participants overestimated peer performance and set targets 20.3% higher than the top performance benchmark. In 68.4% of the cases, intentions to improve practice were consistent with actual gaps in performance (without feedback); it increased to 79.9% after receiving feedback. In 56.3% of the cases, participants still wanted to improve the aspects that they were already top performers in, and in 8.3% of the cases, they lacked improvement intentions, as they did not consider indicators important.

Stattin et al [37] evaluated a SMART (specific, measurable, accepted, realistic, timely) performance feedback dashboard based on data from a national cancer registry. The proportion of patients reported in a timely fashion to the registry increased from 26% in 2011 to 40% in 2013 ( $P<.001$ ). The use of active surveillance for men with very low-risk prostate cancer increased from 63% to 86% ( $P<.001$ ). The waiting time remained long. In 2013, the overall median time from receipt

of referral to the first visit to a specialist clinic was 35 days (IQR 21-58). From prostate biopsy to the date when the patient received information on their cancer diagnosis was 29 days (IQR 21-40).

Weiner et al [22] evaluated a dashboard for leadership to monitor emergency physicians' and radiologist's performance against established targets. They found that acute patients' (who may require admission) monthly length of stay dropped by 54 minutes. Similarly, the monthly length of stay of lower acuity patients (outpatients) dropped by nearly an hour. Finally, the number of patients in the emergency department who left without being seen fell from 165 per month to 10 per month.

Clark et al [34] observed improvements in process indicators during a 3-month intervention of a clinical dashboard that supported decision-making. The indicator performance improved by an average of 21.2% across the 5 indicators (range 8-38). In particular, discharge plans communicated to patients 24 hours before discharge increased from 48% to 86%. In addition, pharmacy scripts written 24 hours before patient discharge increased from 62% to 84%.

Linder et al [21] investigated whether an acute respiratory infection dashboard changed prescription rates. The study found no difference between intervention and control practices in antibiotic prescriptions for all acute respiratory infection visits.

### ***Evaluating User Behavior Through System Access Log Analysis***

Mulhall et al [35] evaluated a dashboard to support primary care physicians in quality improvement. The study found that 50% of the general practitioners viewed the web-based report (n=400), with 90% signing up for email delivery. Participants who viewed at least one of their reports had an almost 2% reduction in antipsychotic prescribing rates.

Patel et al [29] conducted a cluster randomized controlled trial to evaluate a dashboard to support team-based A&F. During the 5-month intervention period, the dashboard was accessed 104 times by 40 users in February, 77 times by 33 users in March, and 55 times by 30 users in April. During the washout period, the dashboard was accessed 48 times by 20 users in May and 48 times by 24 users in June. After a 9-month intervention period, the use logs showed that 28% of clinicians used the dashboard at least once ( $n=72$ ); these clinicians had lower overall acute respiratory infection prescribing rates (42%) compared with the control group (50%;  $P=.02$ ) [21].

### **Expert Usability Method**

Only 6% (1/18) of the studies conducted a heuristic evaluation of the dashboard interface [26]. In all, 2 human factor professionals and 3 focus group members evaluated the dashboard based on Nielsen 10 heuristics [46]. The expert review identified 20 suggestions for the changes. Overall, 5 changes were recommended by 40% or more of the evaluators. The top suggestion (with 60% of evaluators in agreement) was to include a cover sheet documenting the goal of the program and quality indicator criteria (clarity) and to remove the catheter quality indicator (repetitive).

### **User Study Usability Methods**

The key results from the studies that involved end-user usability testing of the interfaces were generally positive. These studies required participants to complete predefined tasks on the interface to identify errors and measure the time to completion.

Janssen et al [33] found that 5 participants completed all the think-aloud protocol tasks on the individual patient dashboard ( $n=5$ ). On the cohort dashboard, only 1 of the 5 participants was able to complete the first task to identify the proportion of patients with lymphedema that had >10 resected nodes. The last task on the cohort dashboard, which required participants to identify the proportion of patients within the organization having ongoing treatment for lymphedema and a BMI in the overweight range, was only attempted by 3 participants.

Brown et al [38] evaluated an e-A&F dashboard to understand the optimal interface design for the clinical A&F process. In all, 7 participants identified a median of 10 errors (range 8-21). A median of 5 tasks were completed out of the 7 evaluation tasks (range 4-7); 16% (6/38) of the possible heuristic categories were violated, with the most frequently violated being *workflow integration* ( $n=40$ ).

Schall et al [24] observed that the time on task improved in 6 of the 8 evaluation tasks between the conventional and HIT dashboards ( $n=6$ ). In terms of accuracy, the tasks completed without errors improved across 5 of the 8 tasks. Task completion without errors remained the same between the conventional and HIT dashboards in the first 2 evaluation tasks. Tasks completed without errors decreased in 1 evaluation task (pressure ulcers).

### **RQ6: Future Design Considerations**

There were 4 key themes that were identified across the included studies related to future dashboard design considerations.

### **Engagement With Clinical Staff**

A key design consideration was the involvement of end users throughout the development process. For example, Laurent et al [23] followed a user-centered process when developing a tool to guarantee usability and ensured that the information displayed did not lead to misunderstandings or interpretation errors.

Promoting dashboards through demonstrations at meetings with individuals or teams was suggested by Schall et al [24] as a strategy to engage clinical staff. To fully integrate the dashboard use in practice, the study suggested updating practice reminders, providing actionable feedback of quality improvement data, and reporting to senior leaders. In addition, local champions or change agents in each unit were responsible for using the dashboard during interprofessional daily huddles.

### **Clinical Indicators**

The selection of clinical indicators was a common topic discussed in the included studies. Stattin et al [37] noted that the selection of quality indicators should be based on recently published guidelines that have been widely accepted. Patel et al [29] highlighted that to be effective in improving care, the use of process indicators that the evidence trying to measure an outcome is continuously evaluated, and providers have the opportunity to provide feedback on how meaningful they find the measures. The principle of fairness should also be considered when selecting clinical indicators; specifically, performance standards need to be evaluated and set concerning quality care, for example, the minimum standards for competency in residency programs [31].

When presenting clinical indicators to clinicians, Brown et al [38] suggested that indicators should be framed positively where appropriate to emphasize achievement. In addition, clinical indicators should be prioritized automatically.

Linder et al [21] highlighted that reporting clinical indicators, by itself, is frequently insufficient to improve the quality of care. Linder et al [21] suggested that quality reporting likely needs to be coupled with other interventions, such as *clinical detailing, clinical decision support, patient education, or financial incentives*. Clark et al [34] also suggested cointerventions, such as a dashboard, including decision-support tools.

Herzke et al [30] highlighted the benefits of attributing performance data to individual clinicians rather than admitting clinicians. However, the authors warned that the computational requirements of their methodology were not trivial and required linking billing data with administrative patient-level data, which may be challenging to operationalize.

Gude et al [36] proposed that more intensive measures, such as verbal feedback and feedback discussions in teams rather than among individuals might be required to ensure clinicians recognize the importance of indicators and trust in data.

### **Support to Interpret Performance Data**

The studies identified clinicians having difficulty interpreting the clinical indicator data to make sense of their individual and team performance. To support clinicians in interpreting their

data, Schall et al [26] suggested that if quality indicator scores do not have meaning, the score should not be included. If included, more precise definitions of symbol color and quality indicators would be helpful. Similarly, a cover sheet documenting the goal of the program and quality indicator criteria was also proposed.

To prevent benchmarks from being perceived as unrealistically high, Gude et al [36] recommended delivering multiple performance comparators, such as median, top 10% peer performance, and own past performance. Ranking individual provider performance relative to peers was also suggested by Herzke et al [30].

Brown et al [38] proposed comparing scores of users to desirable performance labels, such as using a traffic light system to reduce ambiguity. If the dashboard presents suggested actions, it should provide further data analysis and visualization related to recommended improvement actions and clearly explain what performance data specifically refer to.

To address the known issues around attributing performance data between admitting and consulting clinicians during the episode of care, Herzke et al [30] found that ensuring that data can be credibly attributed to the individual provider was integral in dashboard design.

Dashboards should have the ability to provide details on demand related to why particular improvement actions were suggested, how they have been implemented in other organizations, and patient-facing information [38]. For intermittent dashboard use, Janssen et al [33] suggested it may be helpful to add scaffolding to support exploration of key aspects of practice performance and a history mechanism to enable clinicians and administrators to track progress and changes.

### **Technology**

Broader technology considerations were also highlighted in the included studies. Stattin et al [37] described a scenario in which clinicians may not log in to the dashboard platform. Emails should distribute quarterly reports to department heads to support clinicians' adoption of new technology.

Looking into the future of dashboards based on repurposed clinical indicator data, Clark et al [34] outlined the need for dashboards to continue to focus on quality metrics and to include decision-support tools. In addition, Clark et al [34] predict that initiatives that focus on improving patient experience, such as patient-reported satisfaction, will feature on future dashboards incorporating predictive modeling within dashboards to provide a broader set of information for clinicians.

### **Continuing Professional Development**

Activities completed by clinicians involving reviewing their performance and measuring patient health outcomes are considered CPD activities in specialist professional performance frameworks.

For dashboards that include suggested improvement actions, Brown et al [38] suggested allowing clinicians to add their own actions, which should be saved automatically. Clinicians should

also be allowed to easily save, mark actions as implemented, and view those of other users within their organization.

Mulhall et al [35] identified the added benefits of an e-A&F dashboard. These reports can be used as part of a self-reflective study toward continuing medical education credits required in Ontario [35].

## **Discussion**

### **Principal Findings**

#### **Overview**

The results of this scoping review summarized and mapped the existing literature on emerging performance feedback dashboards based on routinely collected clinical indicator data. The scoping review adds to the literature in several ways. First, the review provides an overview of the different contexts in which these interfaces are used. Second, the review identified common visual and functional features. Third, this review summarizes the design processes and evaluation methods. Finally, the review reports the key outcomes of the included studies and the future design considerations proposed by the authors.

The following section discusses the review implications with respect to the initial RQs.

#### **RQ1: Purpose**

##### **The Purpose of the Dashboards Included Performance Improvement, Quality and Safety, and Management of Operations**

Performance or quality improvement interfaces are focused on presenting relevant clinical indicator data to allow clinicians to reflect on their individual and team performance.

##### **There Is Potential to Improve Support for Interpretation**

Dashboards may have scaffolding questions to support a clinician's metacognitive processes and suggest improvement actions to implement. However, only 2 studies [35,38] have used these techniques to support the end users interpret their performance data. Guidance for clinicians to make sense of their performance data was a common theme identified across the included studies.

#### **RQ2 and RQ3: Common Features and Design Processes**

##### **Generic Indicators Dominated the Studies**

Most of the underlying clinical indicators used to populate the dashboards were categorized according to Mainz et al [8] as *process* and *generic* indicators. Length of stay, 28-day readmissions, and late discharges measure the activities in episodes of care. Generic indicators are not only relevant to specific specialties or subspecialties. Generic process indicators seem to be suitable indicators, as most of the dashboards were designed for team use in a specialty craft group or a multidisciplinary team.

##### **Most Dashboards Were Designed for Group Use**

As Herzke et al [30] reported, it is difficult to attribute the performance of individual clinicians when multiple consultants

can interact with a patient during a single episode of care. Therefore, it is important for the whole team responsible for outcomes to see the performance indicators and work together to review that information.

### **The Studies Had a Similar Number of Fast and Slow Dashboards**

The dashboards were evenly split for fast (<1 day) and slow use (weekly, monthly, or quarterly). Dashboards designed for emergency, intensive care, and maternity wards tend to be for fast use, where data should be optimized for constant monitoring and legible at a glance. Slow dashboards emphasized changes in clinical indicators over time, comparisons with peers, and reflection and goal-setting features.

### **Clinicians Were Engaged During the Design Process, but Details Were Often Unclear for Reproducibility**

Although some of the studies in this review used user-centered and co-design approaches, most studies did not provide details on how their interfaces were designed. Without a description of the design approach and methods used, it is difficult for the studies to be replicated in future studies. By conducting user research and involving clinicians in the design process, HIT projects shift the role of designers from being experts to facilitators of the design process [47]. End users, such as clinicians, medical administrators, nurses, and allied health practitioners are empowered to engage in the design process. Ultimately, researchers gain a deeper understanding of the context of end users and create solutions that address real-world problems. Increased staff engagement was evident in the study by Mulhall et al [35], where the authors used co-design methods to develop a dashboard to improve prescribing rates. They observed a 2% reduction in antipsychotic prescriptions, and most of the participants (n=316) stated that they liked the dashboard and were likely to implement suggested practice changes.

### **RQ4 and RQ5: Evaluation Methods and Reported Outcomes**

#### **Dashboards Were Evaluated Either in a Controlled Laboratory Setting or an Authentic in-the-Wild Environment**

Laboratory studies, such as usability testing, allow researchers to identify whether users are able to complete intended tasks on the interface with minimal errors. The use of standardized usability questionnaires allows researchers to compare utility and satisfaction scores among similar studies. On the other hand, in-the-wild studies allow researchers to identify adoption and implementation issues as the intervention is deployed in authentic environments, such as emergency wards and primary care clinics. Researchers are able to identify changes in actual performance by analyzing eHealth data in EHRs and patient administration systems. Analyzing system access log data provides another perspective on user behavior, allowing researchers to compare how participants thought they used the system with their actual use patterns.

### **Overwhelmingly Positive Responses From Participants on Dashboard Usefulness and Ease of Use**

The standardized questionnaires, surveys, and interviews revealed that most participants found the dashboards to be useful and easy to use. Although the 5 studies reported high usability scores [23,24,26,33,38], none of the studies investigated changes in the underlying clinical indicators before and after clinicians used the dashboard in practice. Future dashboard studies should consider conducting a mix of controlled laboratory user testing and in-the-wild studies to understand the users' initial reaction to an interface and its impact on practice.

### **A Majority Showed Promising Results, Even for the Small Set of in-the-Wild Studies That Assessed Improvements to Clinical Indicators**

Only 22% (2/9) of studies reported no significant change in clinical indicators after the intervention period, whereas 78% (7/9) reported improvements. In all, 22% (2/9) of studies [21,29] revealed that <50% of the participants in their study accessed the dashboard during the evaluation period, suggesting issues with implementation and adoption. Although the studies evaluated the dashboards in either a controlled laboratory or authentic in-the-wild environment, no study evaluated the dashboard in both settings. Studies evaluating the impact of dashboards on clinical performance can be improved by incorporating data collected from controlled and authentic settings. Laboratory studies provide insight into the usability (accuracy, efficiency, and satisfaction) of interfaces for target users to achieve specific goals. Authentic in-the-wild studies provide insight into whether the interface supports changes in individual practice, internal processes, and patient health outcomes. Together, these environments provide rich data on the usefulness and effectiveness of the interface.

### **RQ6: Future Design Considerations**

#### **Engaging Clinicians During the Design Process Is Integral in Successful Implementation of Dashboards**

The key themes identified related to future dashboards focus on the importance of engaging with staff in the design process, selecting and presenting appropriate clinical indicators, and supporting clinicians in interpreting performance data. Clinical staff must be involved in the design process of dashboards to support reflection on practice. The design should consider the differences in how care teams work, such as individual and team displays, frequency of use (at-a-glance vs long-term use), and specialty specific clinical indicators. The indicators selected must be relevant to the clinician's work and presented to maximize understanding. Aligning the design of the dashboard interfaces to existing clinical performance improvement frameworks, such as Plan-Do-Study-Act [39], could better support clinicians in interpreting their individual and team performance data. Some techniques that could support this process of sense-making and reflection include scaffolding questions and annotation features with private notes [33].

#### **Scaffolding Reflection to Support Long-term Professional Learning**

The articles largely evaluated the usability and usefulness of the performance feedback interfaces. The current set of articles

in this review suggests that the dashboards were relevant to their practice and that clinicians had a strong understanding of the clinical indicator data presented, both important steps in the reflection process. The interfaces received positive results related to the self-monitoring of performance data. However, the interfaces lacked features designed to support metacognitive processes, such as self-reflection, planning, and goal-setting [48]. One study incorporated features to suggest improvement actions based on guidelines and the ability to save personal improvement actions [38]. There is an opportunity to better understand how clinicians and teams make sense of their performance data, particularly how clinicians conclude that current practice is appropriate, when to initiate change in behavior, or conclude that past practice change initiatives have been effective. Clinicians may undertake improvement actions, such as conducting an A&F project, peer mentoring, or upskilling through the completion of CPD activities.

### Limitations

The review was restricted to specific databases (MEDLINE, Embase, Scopus, ACM Digital Library, and Web of Science), and a defined search query. This search strategy breadth was refined in consultation with the authors and the university librarian to ensure that the search captured articles across the health informatics, human-computer interaction, and data visualization fields.

As a result, the review does not include articles indexed by CINAHL, Cochrane, PsycINFO, and ERIC. The review does not include gray literature because no quality assessment of

studies was planned to be conducted; therefore, the search was restricted to peer-reviewed articles only. Despite care in the design of the search process, some studies may not have been captured owing to their journal or indexing bias.

The search was restricted to a specific time frame to ensure the review was feasible to conduct, which could have led to some older studies being excluded. Limiting the time frame to the last 10 years ensured that the review captured changes in technology. The time frame also focused on state-of-the-art case studies rather than on the history of clinical dashboards.

The review followed the guidelines by Arksey and O'Malley, where a quality assessment of the studies is not required. As such, the review identifies a breadth of research, although this includes work that may not have been validated.

### Conclusions

Our work was motivated by the need for effective tools that support clinicians in reflecting on their practice. This scoping review mapped the current landscape of literature on dashboards based on routinely collected clinical indicator data to support reflection. Although there were common data visualization techniques and clinical indicators used across studies, there was variance in the design and evaluation of the dashboards. We identified a lack of interface features to support clinicians in making sense and reflecting on their personal performance data. We conclude that there is a gap in the literature on dashboards based on routinely collected clinical indicator data that are personalized and scaffolded visualization interfaces to support long-term reflection.

---

### Acknowledgments

The research described in this paper was funded through a project grant from the Digital Health Cooperative Research Centre (DHCRC). The DHCRC was established and supported by the Australian Government's Cooperative Research Centers Program. The DHCRC project primarily grants funding scholarships for higher-degree research candidates and postdoctoral research fellows. Author BB holds a higher degree of research scholarship funded through a DHCRC project grant. The authors would like to thank Jun Guo from the University of Sydney Library for support in the search strategy. The authors would like to thank Carol Pizzuti for screening the articles for this scoping review.

---

### Conflicts of Interest

None declared.

---

#### Multimedia Appendix 1

Example MEDLINE search terms.

[\[DOCX File, 24 KB - medinform\\_v10i2e32695\\_app1.docx\]](#)

---

#### Multimedia Appendix 2

Summary of characteristics of studies included in the scoping review.

[\[DOCX File, 29 KB - medinform\\_v10i2e32695\\_app2.docx\]](#)

---

#### Multimedia Appendix 3

Summary of results of studies included in the scoping review.

[\[DOCX File, 38 KB - medinform\\_v10i2e32695\\_app3.docx\]](#)

---

### References

1. Sarikaya A, Correll M, Bartram L, Tory M, Fisher D. What do we talk about when we talk about dashboards? *IEEE Trans Vis Comput Graph* 2019 Jan;25(1):683. [doi: [10.1109/TVCG.2018.2864903](https://doi.org/10.1109/TVCG.2018.2864903)] [Medline: [30136958](https://pubmed.ncbi.nlm.nih.gov/30136958/)]
2. Khairat SS, Dukkipati A, Lauria HA, Bice T, Travers D, Carson SS. The impact of visualization dashboards on quality of care and clinician satisfaction: integrative literature review. *JMIR Hum Factors* 2018 May 31;5(2):e22 [FREE Full text] [doi: [10.2196/humanfactors.9328](https://doi.org/10.2196/humanfactors.9328)] [Medline: [29853440](https://pubmed.ncbi.nlm.nih.gov/29853440/)]
3. Faiola A, Srinivas P, Duke J. Supporting clinical cognition: a human-centered approach to a novel ICU information visualization dashboard. *AMIA Annu Symp Proc* 2015 Nov 5;2015:560-569 [FREE Full text] [Medline: [26958190](https://pubmed.ncbi.nlm.nih.gov/26958190/)]
4. Brown B, Balatsoukas P, Williams R, Sperrin M, Buchan I. Interface design recommendations for computerised clinical audit and feedback: hybrid usability evidence from a research-led system. *Int J Med Inform* 2016 Oct;94:191-206 [FREE Full text] [doi: [10.1016/j.ijmedinf.2016.07.010](https://doi.org/10.1016/j.ijmedinf.2016.07.010)] [Medline: [27573327](https://pubmed.ncbi.nlm.nih.gov/27573327/)]
5. Gude WT, van Engen-Verheul MM, van der Veer SN, de Keizer NF, Peek N. How does audit and feedback influence intentions of health professionals to improve practice? A laboratory experiment and field study in cardiac rehabilitation. *BMJ Qual Saf* 2017 Apr;26(4):279-287. [doi: [10.1136/bmjqs-2015-004795](https://doi.org/10.1136/bmjqs-2015-004795)] [Medline: [27068999](https://pubmed.ncbi.nlm.nih.gov/27068999/)]
6. Freeman T. Using performance indicators to improve health care quality in the public sector: a review of the literature. *Health Serv Manage Res* 2002 May;15(2):126-137. [doi: [10.1258/0951484021912897](https://doi.org/10.1258/0951484021912897)] [Medline: [12028801](https://pubmed.ncbi.nlm.nih.gov/12028801/)]
7. Shaw T, Janssen A, Crampton R, O'Leary F, Hoyle P, Jones A, et al. Attitudes of health professionals to using routinely collected clinical data for performance feedback and personalised professional development. *Med J Aust* 2019 Apr;210 Suppl 6:S17-S21. [doi: [10.5694/mja2.50022](https://doi.org/10.5694/mja2.50022)] [Medline: [30927464](https://pubmed.ncbi.nlm.nih.gov/30927464/)]
8. Mainz J. Defining and classifying clinical indicators for quality improvement. *Int J Qual Health Care* 2003 Dec;15(6):523-530. [doi: [10.1093/intqhc/mzg081](https://doi.org/10.1093/intqhc/mzg081)] [Medline: [14660535](https://pubmed.ncbi.nlm.nih.gov/14660535/)]
9. Davis N, Davis D, Bloch R. Continuing medical education: AMEE Education Guide No 35. *Med Teach* 2008;30(7):652-666. [doi: [10.1080/01421590802108323](https://doi.org/10.1080/01421590802108323)] [Medline: [18777424](https://pubmed.ncbi.nlm.nih.gov/18777424/)]
10. Kaufman D. Applying educational theory in practice. *BMJ* 2003 Jan 25;326(7382):213-216 [FREE Full text] [doi: [10.1136/bmj.326.7382.213](https://doi.org/10.1136/bmj.326.7382.213)] [Medline: [12543841](https://pubmed.ncbi.nlm.nih.gov/12543841/)]
11. 2021 MyCPD framework. The Royal Australasian College of Physicians. URL: <https://www.racp.edu.au/fellows/continuing-professional-development/2021-mycpd-framework> [accessed 2021-12-26]
12. Continuing professional development. Royal Australasian College of Surgeons. URL: <https://www.surgeons.org/en/Fellows/continuing-professional-development> [accessed 2021-12-26]
13. Ivers N, Jamtvedt G, Flottorp S, Young J, Odgaard-Jensen J, French S, et al. Audit and feedback: effects on professional practice and healthcare outcomes. *Cochrane Database Syst Rev* 2012 Jun 13(6):CD000259. [doi: [10.1002/14651858.CD000259.pub3](https://doi.org/10.1002/14651858.CD000259.pub3)] [Medline: [22696318](https://pubmed.ncbi.nlm.nih.gov/22696318/)]
14. Ivers NM, Grimshaw JM, Jamtvedt G, Flottorp S, O'Brien MA, French SD, et al. Growing literature, stagnant science? Systematic review, meta-regression and cumulative analysis of audit and feedback interventions in health care. *J Gen Intern Med* 2014 Nov;29(11):1534-1541 [FREE Full text] [doi: [10.1007/s11606-014-2913-y](https://doi.org/10.1007/s11606-014-2913-y)] [Medline: [24965281](https://pubmed.ncbi.nlm.nih.gov/24965281/)]
15. Tuti T, Nzinga J, Njoroge M, Brown B, Peek N, English M, et al. A systematic review of electronic audit and feedback: intervention effectiveness and use of behaviour change theory. *Implement Sci* 2017 May 12;12(1):61 [FREE Full text] [doi: [10.1186/s13012-017-0590-z](https://doi.org/10.1186/s13012-017-0590-z)] [Medline: [28494799](https://pubmed.ncbi.nlm.nih.gov/28494799/)]
16. van Deen WK, Cho ES, Pustolski K, Wixon D, Lamb S, Valente TW, et al. Involving end-users in the design of an audit and feedback intervention in the emergency department setting - a mixed methods study. *BMC Health Serv Res* 2019 Apr 29;19(1):270 [FREE Full text] [doi: [10.1186/s12913-019-4084-3](https://doi.org/10.1186/s12913-019-4084-3)] [Medline: [31035992](https://pubmed.ncbi.nlm.nih.gov/31035992/)]
17. Arksey H, O'Malley L. Scoping studies: towards a methodological framework. *Int J Soc Res Method* 2005 Feb;8(1):19-32. [doi: [10.1080/1364557032000119616](https://doi.org/10.1080/1364557032000119616)]
18. MacEntee MI. A typology of systematic reviews for synthesising evidence on health care. *Gerodontology* 2019 Dec;36(4):303-312. [doi: [10.1111/ger.12439](https://doi.org/10.1111/ger.12439)] [Medline: [31691365](https://pubmed.ncbi.nlm.nih.gov/31691365/)]
19. Peters MD, Godfrey CM, Khalil H, McInerney P, Parker D, Soares CB. Guidance for conducting systematic scoping reviews. *Int J Evid Based Healthc* 2015 Sep;13(3):141-146. [doi: [10.1097/XEB.000000000000050](https://doi.org/10.1097/XEB.000000000000050)] [Medline: [26134548](https://pubmed.ncbi.nlm.nih.gov/26134548/)]
20. Kay J, Rus V, Zapata-Rivera D, Durlach P. Open Learner Model Visualisations for Contexts Where Learners Think Fast or Slow. In: *Design Recommendations for Intelligent Tutoring Systems*. Orlando, Florida: US Army Combat Capabilities Development Command - Soldier Center; Dec 30, 2020:126.
21. Linder J, Schnipper J, Tsurikova R, Yu T, Volk L, Melnikas A, et al. Electronic health record feedback to improve antibiotic prescribing for acute respiratory infections. *Am J Manag Care* 2010 Dec;16(12 Suppl HIT):e311-e319. [Medline: [21322301](https://pubmed.ncbi.nlm.nih.gov/21322301/)]
22. Weiner J, Balijepally V, Tanniru M. Integrating strategic and operational decision making using data-driven dashboards: the case of St. Joseph mercy Oakland hospital. *J Healthc Manag* 2015;60(5):319-330. [Medline: [26554142](https://pubmed.ncbi.nlm.nih.gov/26554142/)]
23. Laurent G, Moussa MD, Cirenei C, Tavernier B, Marcilly R, Lamer A. Development, implementation and preliminary evaluation of clinical dashboards in a department of anesthesia. *J Clin Monit Comput* 2021 May;35(3):617-626 [FREE Full text] [doi: [10.1007/s10877-020-00522-x](https://doi.org/10.1007/s10877-020-00522-x)] [Medline: [32418147](https://pubmed.ncbi.nlm.nih.gov/32418147/)]
24. Schall M, Cullen L, Pennathur P, Chen H, Burrell K, Matthews G. Usability evaluation and implementation of a health information technology dashboard of evidence-based quality indicators. *Comput Inform Nurs* 2017 Jun;35(6):281-288. [doi: [10.1097/CIN.0000000000000325](https://doi.org/10.1097/CIN.0000000000000325)] [Medline: [28005564](https://pubmed.ncbi.nlm.nih.gov/28005564/)]



25. Patel M, Rathi B, Tashfeen K, Yarubi MA. Development and implementation of maternity dashboard in regional hospital for quality improvement at ground level: a pilot study. *Oman Med J* 2019 May;34(3):194-199 [FREE Full text] [doi: [10.5001/omj.2019.38](https://doi.org/10.5001/omj.2019.38)] [Medline: [31110625](https://pubmed.ncbi.nlm.nih.gov/31110625/)]
26. Schall MC, Chen H, Pennathur PR, Cullen L. Development and evaluation of a health information technology dashboard of quality indicators. *Proc Human Fact Ergonom Soc Annual Meet* 2016 Dec 20;59(1):461-465. [doi: [10.1177/1541931215591099](https://doi.org/10.1177/1541931215591099)]
27. Hester G, Lang T, Madsen L, Tambyraja R, Zenker P. Timely data for targeted quality improvement interventions: use of a visual analytics dashboard for bronchiolitis. *Appl Clin Inform* 2019 Jan;10(1):168-174 [FREE Full text] [doi: [10.1055/s-0039-1679868](https://doi.org/10.1055/s-0039-1679868)] [Medline: [30841007](https://pubmed.ncbi.nlm.nih.gov/30841007/)]
28. Khanna N, Gritzer L, Klyushnenkova E, Montgomery R, Dark M, Shah S, et al. Practice transformation analytics dashboard for clinician engagement. *Ann Fam Med* 2019 Aug 12;17(Suppl 1):S73-S76 [FREE Full text] [doi: [10.1370/afm.2382](https://doi.org/10.1370/afm.2382)] [Medline: [31405880](https://pubmed.ncbi.nlm.nih.gov/31405880/)]
29. Patel S, Rajkomar A, Harrison JD, Prasad PA, Valencia V, Ranji SR, et al. Next-generation audit and feedback for inpatient quality improvement using electronic health record data: a cluster randomised controlled trial. *BMJ Qual Saf* 2018 Sep;27(9):691-699. [doi: [10.1136/bmjqs-2017-007393](https://doi.org/10.1136/bmjqs-2017-007393)] [Medline: [29507124](https://pubmed.ncbi.nlm.nih.gov/29507124/)]
30. Herzke CA, Michtalik HJ, Durkin N, Finkelstein J, Deutschendorf A, Miller J, et al. A method for attributing patient-level metrics to rotating providers in an inpatient setting. *J Hosp Med* 2018 Jul 01;13(7):470-475. [doi: [10.12788/jhm.2897](https://doi.org/10.12788/jhm.2897)] [Medline: [29261820](https://pubmed.ncbi.nlm.nih.gov/29261820/)]
31. Ehrenfeld J, McEvoy M, Furman W, Snyder D, Sandberg W. Automated near-real-time clinical performance feedback for anesthesiology residents: one piece of the milestones puzzle. *Anesthesiology* 2014 Jan;120(1):172-184 [FREE Full text] [doi: [10.1097/ALN.0000000000000071](https://doi.org/10.1097/ALN.0000000000000071)] [Medline: [24398735](https://pubmed.ncbi.nlm.nih.gov/24398735/)]
32. Leahy I, Borczuk R, Ferrari L. Making the pediatric perioperative surgical home come to life by leveraging existing health information technology. *Curr Opin Anaesthesiol* 2017 Jun;30(3):383-389. [doi: [10.1097/ACO.0000000000000454](https://doi.org/10.1097/ACO.0000000000000454)] [Medline: [28291130](https://pubmed.ncbi.nlm.nih.gov/28291130/)]
33. Janssen A, Donnelly C, Kay J, Thiem P, Saavedra A, Pathmanathan N, et al. Developing an intranet-based lymphedema dashboard for breast cancer multidisciplinary teams: design research study. *J Med Internet Res* 2020 Apr 21;22(4):e13188 [FREE Full text] [doi: [10.2196/13188](https://doi.org/10.2196/13188)] [Medline: [32314968](https://pubmed.ncbi.nlm.nih.gov/32314968/)]
34. Clark KW, Whiting E, Rowland J, Thompson LE, Missenden I, Schellein G. Breaking the mould without breaking the system: the development and pilot of a clinical dashboard at The Prince Charles Hospital. *Aust Health Rev* 2013 Jun;37(3):304-308. [doi: [10.1071/AH12018](https://doi.org/10.1071/AH12018)] [Medline: [23701819](https://pubmed.ncbi.nlm.nih.gov/23701819/)]
35. Mulhall CL, Lam JM, Rich PS, Dobell LG, Greenberg A. Enhancing quality care in Ontario long-term care homes through audit and feedback for physicians. *J Am Med Dir Assoc* 2020 Mar;21(3):420-425. [doi: [10.1016/j.jamda.2019.11.017](https://doi.org/10.1016/j.jamda.2019.11.017)] [Medline: [31974064](https://pubmed.ncbi.nlm.nih.gov/31974064/)]
36. Gude WT, Roos-Blom M, van der Veer SN, Dongelmans DA, de Jonge E, Francis JJ, et al. Health professionals' perceptions about their clinical performance and the influence of audit and feedback on their intentions to improve practice: a theory-based study in Dutch intensive care units. *Implement Sci* 2018 Feb 17;13(1):33 [FREE Full text] [doi: [10.1186/s13012-018-0727-8](https://doi.org/10.1186/s13012-018-0727-8)] [Medline: [29454393](https://pubmed.ncbi.nlm.nih.gov/29454393/)]
37. Stattin P, Sandin F, Sandbäck T, Damber J, Franck Lissbrant I, Robinson D, et al. Dashboard report on performance on select quality indicators to cancer care providers. *Scand J Urol* 2016;50(1):21-28. [doi: [10.3109/21681805.2015.1063083](https://doi.org/10.3109/21681805.2015.1063083)] [Medline: [26162012](https://pubmed.ncbi.nlm.nih.gov/26162012/)]
38. Brown B, Balatsoukas P, Williams R, Sperrin M, Buchan I. Multi-method laboratory user evaluation of an actionable clinical performance information system: implications for usability and patient safety. *J Biomed Inform* 2018 Jan;77:62-80 [FREE Full text] [doi: [10.1016/j.jbi.2017.11.008](https://doi.org/10.1016/j.jbi.2017.11.008)] [Medline: [29146562](https://pubmed.ncbi.nlm.nih.gov/29146562/)]
39. Moen R, Norman C. Circling back. *Quality Progress*. 2010. URL: <https://deming.org/wp-content/uploads/2020/06/circling-back.pdf> [accessed 2021-12-26]
40. Shneiderman B, Plaisant C, Hesse BW. Improving healthcare with interactive visualization. *Comput* 2013 May;46(5):58-66. [doi: [10.1109/MC.2013.38](https://doi.org/10.1109/MC.2013.38)]
41. Shneiderman B. The eyes have it: a task by data type taxonomy for information visualizations. In: *Proceedings of the 1996 IEEE Symposium on Visual Languages*. 1996 Presented at: Proceedings 1996 IEEE Symposium on Visual Languages; Sep 3-6, 1996; Boulder, CO, USA. [doi: [10.1109/vl.1996.545307](https://doi.org/10.1109/vl.1996.545307)]
42. Brooke J. SUS: a 'quick and dirty' usability scale. In: *Usability Evaluation In Industry*. Boca Raton, Florida: CRC Press; 1996:191.
43. Bangor A, Kortum PT, Miller JT. An empirical evaluation of the system usability scale. *Int J Human Comput Interact* 2008 Jul 30;24(6):574-594. [doi: [10.1080/10447310802205776](https://doi.org/10.1080/10447310802205776)]
44. Lewis JR. Psychometric evaluation of the PSSUQ using data from five years of usability studies. *Int J Human Comput Interact* 2002 Sep;14(3-4):463-488. [doi: [10.1080/10447318.2002.9669130](https://doi.org/10.1080/10447318.2002.9669130)]
45. Sauro J, Dumas J. Comparison of three one-question, post-task usability questionnaires. In: *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. 2009 Presented at: SIGCHI Conference on Human Factors in Computing Systems; Apr 4 - 9, 2009; Boston MA USA. [doi: [10.1145/1518701.1518946](https://doi.org/10.1145/1518701.1518946)]

46. Nielsen J. Usability inspection methods. In: Proceedings of the Conference Companion on Human Factors in Computing Systems. 1994 Presented at: Conference Companion on Human Factors in Computing Systems; Apr 24 - 28, 1994; Boston Massachusetts USA. [doi: [10.1145/259963.260531](https://doi.org/10.1145/259963.260531)]
47. Maguire M. Methods to support human-centred design. *Int J Human Comput Stud* 2001 Oct;55(4):587-634. [doi: [10.1006/ijhc.2001.0503](https://doi.org/10.1006/ijhc.2001.0503)]
48. Matcha W, Uzir NA, Gasevic D, Pardo A. A systematic review of empirical studies on learning analytics dashboards: a self-regulated learning perspective. *IEEE Trans Learn Technol* 2020 Apr 1;13(2):226-245. [doi: [10.1109/tlt.2019.2916802](https://doi.org/10.1109/tlt.2019.2916802)]

## Abbreviations

**A&F:** audit and feedback  
**CPD:** continuing professional development  
**DHCRC:** Digital Health Cooperative Research Centre  
**e-A&F:** electronic audit and feedback  
**EHR:** electronic health record  
**EMR:** electronic medical record  
**HIT:** health care information technology  
**PSSUQ:** Post-Study System Usability Questionnaire  
**RQ:** research question  
**SEQ:** single ease question  
**SMART:** specific, measurable, accepted, realistic, timely  
**SUS:** system usability scale

*Edited by C Lovis; submitted 06.08.21; peer-reviewed by P Gough, A Lamer; comments to author 24.10.21; revised version received 19.11.21; accepted 04.12.21; published 14.02.22.*

*Please cite as:*

*Bucalon B, Shaw T, Brown K, Kay J*

*State-of-the-art Dashboards on Clinical Indicator Data to Support Reflection on Practice: Scoping Review*

*JMIR Med Inform* 2022;10(2):e32695

URL: <https://medinform.jmir.org/2022/2/e32695>

doi: [10.2196/32695](https://doi.org/10.2196/32695)

PMID: [35156928](https://pubmed.ncbi.nlm.nih.gov/35156928/)

©Bernard Bucalon, Tim Shaw, Kerri Brown, Judy Kay. Originally published in JMIR Medical Informatics (<https://medinform.jmir.org>), 14.02.2022. This is an open-access article distributed under the terms of the Creative Commons Attribution License (<https://creativecommons.org/licenses/by/4.0/>), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Medical Informatics, is properly cited. The complete bibliographic information, a link to the original publication on <https://medinform.jmir.org/>, as well as this copyright and license information must be included.

Review

# The Science of Learning Health Systems: Scoping Review of Empirical Research

Louise A Ellis<sup>1</sup>, PhD; Mitchell Sarkies<sup>1</sup>, PhD; Kate Churruca<sup>1</sup>, PhD; Genevieve Dammery<sup>1</sup>, BSc (Hons); Isabelle Meulenbroeks<sup>1</sup>, MRes;Carolynn L Smith<sup>1</sup>, PhD; Chiara Pomare<sup>1</sup>, PhD; Zeyad Mahmoud<sup>1</sup>, PhD; Yvonne Zurynski<sup>1</sup>, PhD; Jeffrey Braithwaite<sup>1</sup>, PhD

Australian Institute of Health Innovation, Macquarie University, Sydney, Australia

**Corresponding Author:**

Louise A Ellis, PhD

Australian Institute of Health Innovation

Macquarie University

75 Talavera Rd

Sydney, 2113

Australia

Phone: 61 298502484

Fax: 61 298502499

Email: [louise.ellis@mq.edu.au](mailto:louise.ellis@mq.edu.au)

**Related Article:**

This is a corrected version. See correction statement: <https://medinform.jmir.org/2022/8/e41424>

## Abstract

**Background:** The development and adoption of a learning health system (LHS) has been proposed as a means to address key challenges facing current and future health care systems. The first review of the LHS literature was conducted 5 years ago, identifying only a small number of published papers that had empirically examined the implementation or testing of an LHS. It is timely to look more closely at the published empirical research and to ask the question, Where are we now? 5 years on from that early LHS review.

**Objective:** This study performed a scoping review of empirical research within the LHS domain. Taking an “implementation science” lens, the review aims to map out the empirical research that has been conducted to date, identify limitations, and identify future directions for the field.

**Methods:** Two academic databases (PubMed and Scopus) were searched using the terms “learning health\* system\*” for papers published between January 1, 2016, to January 31, 2021, that had an explicit empirical focus on LHSs. Study information was extracted relevant to the review objective, including each study’s publication details; primary concern or focus; context; design; data type; implementation framework, model, or theory used; and implementation determinants or outcomes examined.

**Results:** A total of 76 studies were included in this review. Over two-thirds of the studies were concerned with implementing a particular program, system, or platform (53/76, 69.7%) designed to contribute to achieving an LHS. Most of these studies focused on a particular clinical context or patient population (37/53, 69.8%), with far fewer studies focusing on whole hospital systems (4/53, 7.5%) or on other broad health care systems encompassing multiple facilities (12/53, 22.6%). Over two-thirds of the program-specific studies utilized quantitative methods (37/53, 69.8%), with a smaller number utilizing qualitative methods (10/53, 18.9%) or mixed-methods designs (6/53, 11.3%). The remaining 23 studies were classified into 1 of 3 key areas: ethics, policies, and governance (10/76, 13.2%); stakeholder perspectives of LHSs (5/76, 6.6%); or LHS-specific research strategies and tools (8/76, 10.5%). Overall, relatively few studies were identified that incorporated an implementation science framework.

**Conclusions:** Although there has been considerable growth in empirical applications of LHSs within the past 5 years, paralleling the recent emergence of LHS-specific research strategies and tools, there are few high-quality studies. Comprehensive reporting of implementation and evaluation efforts is an important step to moving the LHS field forward. In particular, the routine use of implementation determinant and outcome frameworks will improve the assessment and reporting of barriers, enablers, and implementation outcomes in this field and will enable comparison and identification of trends across studies.

**KEYWORDS**

learning health systems; learning health care systems; implementation science; evaluation; health system; health care system; empirical research; medical informatics; review

## Introduction

### Background

Contemporary health care systems are not always fit for purpose or evidence-based [1,2]. Despite all the resourcefulness and efforts internationally, health care performance has, by and large, flatlined, with persisting iatrogenic harm, inefficiencies, and health care waste [2,3]. To overcome ongoing challenges in health care systems, there is growing awareness of the need for health care systems predicated on knowledge harvesting and exploitation, and continuing improvement through leveraging big data and incorporating patients' perspectives and choices into decisions [2,4]. The concept of a learning health system (LHS) was first formally discussed at a Roundtable on Evidence-Based Medicine in 2007 [5]. There is now widespread recognition that what is needed is a health care system that "consistently delivers reliable performance and constantly improves, systematically and seamlessly, with each care experience—in short, a system with an ability to learn" [6].

### The Vision for, and Progress Toward, an LHS

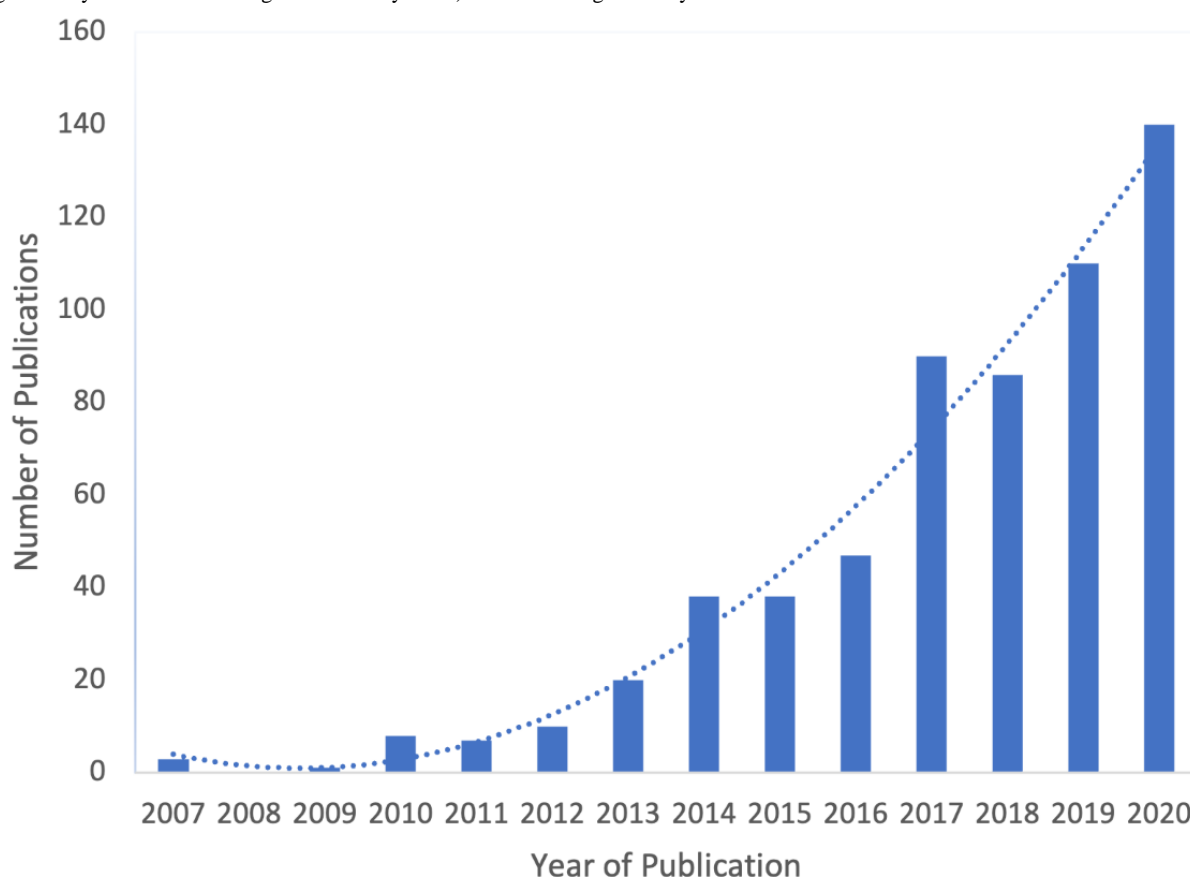
An LHS has been described by the US Institute of Medicine (IoM; now the National Academy of Medicine) as one where science, informatics, incentives, and culture are aligned for enduring continuous improvement and innovation; best practices are seamlessly embedded in the care process; patients and families are active participants in all elements; and new knowledge is captured as an integral by-product of the care experience [7]. Priorities for achieving this lofty, aspirational vision include advancing the development of a fully interoperable digital infrastructure, the application of data-driven research within health care, and a culture of transparency on outcomes and cost [8]. Although this vision has remained largely aspirational to date, rapid innovations in big data, machine learning, and artificial intelligence (AI) are creating the opportunity, and expectation, that health care systems can make real the promise of an LHS [4,9,10]. For example, in the United

States, well-regarded health care provider Geisinger reported on its significantly expanded informatics and science capabilities over the past 5 years by migrating its comprehensive data assets into a big data enterprise data warehouse infrastructure [11]. Geisinger documented its efforts to improve patient-clinician engagement with patient-reported experience measures (PREMS) serving as the primary metric for measuring success, moving Geisinger into closer alignment with the LHS vision [11].

### Empiricizing LHSs

Despite enthusiasm for big data and AI as the learning cornerstones, the question remains whether there is compelling evidence for the successful implementation of programs, systems, and services that are making marked progress toward approximating the normative descriptions of the LHS. Research interest in LHS concepts and ideas has been increasing, as evidenced by the growing number of publications on LHS since it was first discussed in 2007 (Figure 1) and the emergence of the influential journal *Learning Health Systems* [12]. Several reviews of the topic have also now emerged, identifying limited but growing empirical LHS applications. In 2016, Budrionis and Bellika [13] conducted a systematic review of the LHS literature, revealing that of the 32 identified papers, only 13 (40.6%) empirically examined the implementation or testing of an LHS. They also found that of the empirical evaluations, most suffered from substantial methodological and data limitations. Two years later, in 2018, Platt et al [14] conducted a scoping review, showing that although most of the research was theoretical, there was a growing number of empirical publications within the LHS domain [14]. More recently, Enticott et al [15] identified 23 LHS environments internationally; most were enabled by digital data gathered by electronic health records. However, these initiatives were largely identified from gray-literature sources (reports and policies) that were not designed as robust studies to create quality research evidence [15].

**Figure 1.** Increase in publications on LHS over time, 2007-2020 (generated using data from PubMed on publications returned using the search term “learning health system” OR “learning health care system”). LHS: learning health system.



With the growth in empirical contributions in the LHS field, it is timely to examine the published empirical research and to determine the status of the field, 5 years on from the first LHS review of Budrionis and Bellika [13]. For this review, we defined an empirical study as one that reports primary or secondary data gathered by means of a specific methodological approach [16]. We seek to leverage recent developments from the field of implementation science, which aligns closely with a core goal of LHSs, to get more evidence into practice, and to satisfy requirements for continuous quality improvement [17-19].

### This Study

In this paper, we report on a scoping review of empirical research within the LHS domain. We map out the empirical research that has been conducted to date, identify limitations, and identify future directions for the field. The scoping review was designed to answer questions in 3 key research areas:

- What types of empirical contributions within the LHS domain have been conducted?
- What have been the key areas of research?
- What study designs and research methods have been used?

Among the empirical studies examining implementation:

- What implementation outcomes have been examined and what implementation determinants have been identified?
- Which implementation science frameworks and tools have been used?

- What are the current knowledge gaps and methodological limitations of empirical research in the LHS field?

## Methods

### Study Design

Our scoping review followed a protocol that was developed in accordance with the Preferred Reporting Items of Systematic Review and Meta-Analyses Extension for Scoping Reviews (PRISMA-ScR) [20]. A scoping review method, which examines the extent, range, and nature of empirical work on a topic, was used to identify gaps and provide suggestions to improve future empirical research on LHSs [21]. For this review, which focused further on the implementation of an LHS, implementation determinants were defined as barriers and enablers that may prevent or facilitate, respectively, improvements in practice [22], as reported in the included studies. The implementation outcomes taxonomy by Proctor et al [23] was used as a systematic framework for examining implementation-focused LHS studies (ie, acceptability, adoption, appropriateness, feasibility, fidelity, implementation cost, penetration, and sustainability), distinguishing these from service and patient outcomes.

### Search Strategy

A search strategy was developed by the research team and executed in January 2021. Two academic databases (PubMed and Scopus) were searched from January 1, 2016, to January 31, 2021, using the term “learning health\* system\*”.

## Inclusion and Exclusion Criteria

Papers were included if they were (1) published from January 1, 2016, to January 31, 2021, (2) had an explicit focus on LHSs, and (3) were empirical studies. Studies reporting primary or secondary data were considered empirical so long as they provided sufficient information about their methodological approach [16,24]. Peer-reviewed journal articles, peer-reviewed full conferences papers, and book chapters that provided sufficient information about their methodological approach and results were also included. Study protocols, review papers, journal commentaries, and editorials were excluded. Studies not in the English language and not explicitly about LHSs (eg, only used the term in the abstract or conclusion) were also excluded.

## Eligibility Screening

Reference details (including abstracts) were downloaded into the reference management software Endnote X9 [25]. The review team (authors LAE, MS, CP, ZM, and IM) screened the full-text publications to determine their inclusion against criteria, and 5% of the retrieved publications were independently screened by the entire review team to ensure consistent inclusion. Any discrepancies among reviewers' judgements were reviewed by 2 authors (LAE and MS) in consultation with authors YZ and JB.

## Data Extraction

Relevant information was extracted at the full-text review stage using a purpose-designed workbook in Microsoft Excel 365 and included (1) publication details (paper title, year, country of residence of corresponding author, paper type, and paper keywords); (2) primary study focus (thematically coded after data extraction); (3) study context (clinical, hospital, health care system); (4) study design (quantitative, qualitative, mixed methods); (5) study data type (primary or secondary); (6) implementation framework, model, or theory used; and (7) implementation determinants or outcomes examined.

## Assessment of Evidence Quality

Consistent with the LHS review by Enticott et al [26], the Grading of Recommendations Assessment, Development and Evaluation (GRADE) approach was applied to assess the overall

quality of evidence based on the study design [26]. Using the GRADE approach, randomized trials without important limitations provide high-quality evidence, while observational studies without special strengths (eg, the use of an implementation science framework) or with important limitations provide low-quality evidence. GRADE recommends that design factors, such as concurrent controls, can improve the quality of evidence; therefore, studies with concurrent controls without important limitations were assessed as providing medium-quality evidence.

## Data Synthesis and Analysis

Papers were grouped together based on extracted data (eg, study design) and summarized through narrative techniques. The country of the corresponding author was coded by income classification based on World Bank definitions of the gross national income per capita. The 3 categories were low (<US \$1045), middle (US \$1046–\$12,695), and high (>US \$12,696) income [27].

Overarching topic areas were identified through an inductive analysis of publication keywords by 2 authors (LAE and CP). These were extracted by the research team and then cleaned and checked for consistency. During data cleaning, keywords were consolidated in the case of plurals (eg, “intervention” vs “interventions”); however, keywords were kept independent in the case of arguably consistent meaning but different phrasing (eg, “learning health care system” vs “learning health system”) in order to represent the variability of terms used in the LHS field. The keyword data was analyzed for frequency and co-occurrence and graphically presented using Gephi version 0.9.2.

The primary study concern or focus was inductively classified by 2 authors (LAE and MS) into 1 of 4 classifications: (1) specific programs, systems, and platforms; (2) ethics, policies, and governance; (3) stakeholder perspectives of LHSs; and (4) LHS-specific research strategies and tools. Studies that examined implementation outcomes were further reviewed and classified by the 2 authors (LAE and MS) according to 8 implementation outcome categories [23], distinguishing these from service and client outcomes, and with definitions tailored to suit the LHS context (Table 1).

**Table 1.** Definition of implementation outcomes [23].

Domain	Definition
Adoption	Uptake of the LHS <sup>a</sup> initiative by health care professionals and health services
Acceptability	Health care professionals' satisfaction with various aspects of the LHS initiative (eg, content, complexity, comfort, delivery, and credibility)
Appropriateness	Fit, relevance, compatibility, suitability, usefulness, and practicability perceived by health care professionals and patients
Feasibility	The actual fit, utility, and practicability of the program within a health service setting and its subsystems, as reported by health care professionals and managers
Fidelity	The LHS initiative delivered, as intended; adherence by health care professionals; and quality of program delivery
Cost	Financial impact of LHS implementation to the health service or organization
Penetration	Spread or reach of the LHS initiative assessed at the organization or setting level
Sustainability	The extent to which the LHS program is maintained or institutionalized within a health service's standard operations

<sup>a</sup>LHS: learning health system.

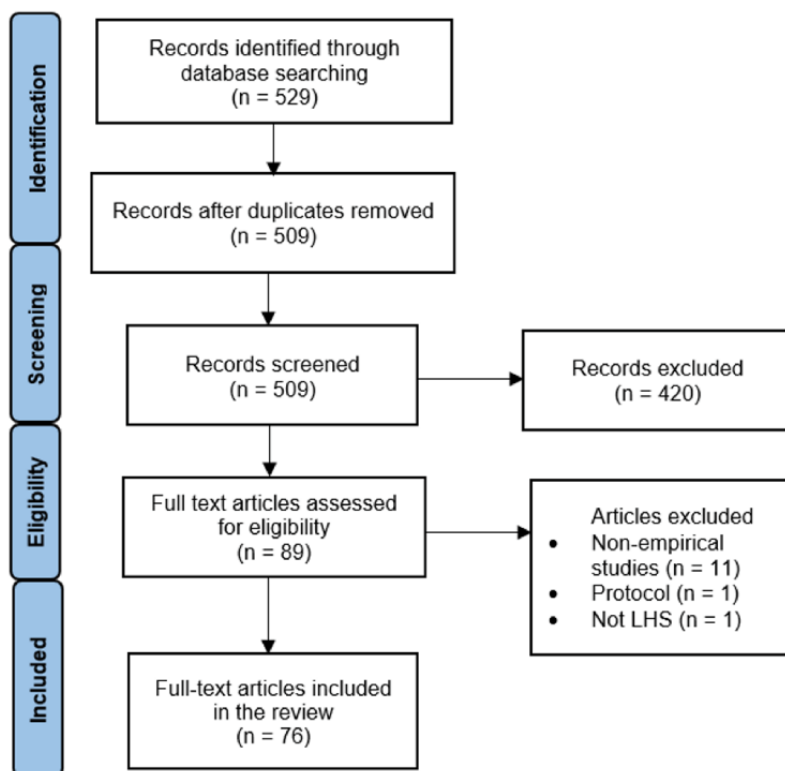
## Results

### Description of Included Studies

The search identified a total of 529 citations. After removing duplicates, 509 (96.2%) remained for title/abstract review. During the title/abstract screening, 420 (82.5%) studies were discarded as not meeting the inclusion criteria. Based on the full-text assessment, a further 13 (14.6%) of 89 studies did not meet the inclusion criteria, and hence 76 (85.4%) studies were included in this review (Figure 2).

A summary of the key characteristics of the included studies is presented in Table 2 (also see Multimedia Appendix 1 for details of all included studies). Of the 76 included studies, the majority (n=72, 94.8%) were published in peer-reviewed journals, 3 (3.9%) were full conference papers, and 1 (1.3%) was a book chapter. The 72 papers were spread widely across 54 different journals, with *Learning Health Systems* (n=7, 9.7%) and *eGEMS (Generating Evidence & Methods to Improve Patient Outcomes)*; n=4, 5.6%) being the most popular.

**Figure 2.** Search and review strategy. LHS: learning health system.



**Table 2.** Summary of key characteristics of the included publications (N=76).

Classification	Papers, n (%)
<b>Country of corresponding author</b>	
United States	55 (72.4)
United Kingdom	9 (11.8)
Canada	2 (7.6)
France	2 (2.6)
Germany	2 (2.6)
The Netherlands	2 (2.6)
Other	4 (5.3)
<b>Country income classification</b>	
High	75 (98.7)
Middle	1 (1.3)
Low	0
<b>Study methods</b>	
Quantitative methods	42 (55.3)
Qualitative methods	27 (35.5)
Mixed methods	7 (9.2)
<b>Study data type</b>	
Primary data	46 (60.5)
Secondary data	23 (30.3)
Both primary and secondary data	7 (9.2)

The location of studies was predominantly restricted to high-income countries, with most coming from the United States (n=55, 72.4%), followed by the United Kingdom (n=9, 11.8%), and Canada (n=3, 3.9%). Over half of the studies (n=42, 55.3%) were quantitative studies, around one-third (n=27, 35.5%) were qualitative and the remaining (n=7, 9.2%) were mixed-methods studies. Although most studies (n=46, 60.5%) utilized primary data alone, one-third of the studies (n=23, 30.3%) relied on secondary data sets, such as electronic health records and data repositories, and a smaller number (n=7, 9.2%) used both primary and secondary data sets.

Of the 76 included studies, 52 (68.4%) listed keywords, with 190 keywords in total. The most common keywords (ie, those with the highest in-degree score) were “learning health system” (n=20), “electronic health records” (n=11), and “learning health care system” (n=8). A network of keywords was created to

demonstrate keywords frequently used in papers together; [Figure 3](#) visually summarizes overarching topic areas of the empirical papers on LHSs identified in this review. To aid interpretation, only keywords with in-degree scores (ie, number of ties directed to or received by a node) greater than or equal to 2 are displayed (see [Figure 3](#)). The size of the node is indicative of frequency (larger nodes indicate a higher number of papers using the keyword). The line between 2 nodes (tie) indicates keywords used together in a paper. Our inductive categorization of keywords identified 4 broad topic areas into which these words fell: (1) study design/methods (eg, comparative effectiveness research, clinical trial, qualitative research), (2) study field (eg, health services research, implementation science), (3) data source (eg, electronic health records), (4) study goal (eg, quality improvement), and (5) barriers/challenges (eg, ethics, data quality). In [Figure 3](#), these are colored separately to indicate keyword categories in relation to one another.



**Figure 3.** Network of co-occurring keywords with in-degree  $\geq 2$ . Each circle (node) is a keyword, and each line (edge) represents co-occurrence. The size of the node indicates the number of times a keyword was used. Colors represent different topic areas. VHA: Veterans Health Administration.



The thematic analysis of the study focus led to classification into either (1) specific programs, systems, and platforms or 1 of the following key research areas: (2) ethics, policies, and governance; (3) stakeholder perspectives of LHSs; or (4) LHS-specific research strategies and tools. This classification system was used to break down studies and separately synthesize information on the study field, setting, population, and study design, as well as implementation determinants and outcomes examined. These categories of focus are considered separately later.

## LHS Programs, Systems, and Platforms

### Implementation and Validation Issues

Over two-thirds of the included studies (53/76, 69.7%) were concerned with implementing a particular program, system, or platform designed to contribute to achieving an LHS. For example, Bhandari et al [28] described the application of a national health outcomes information registry for pain medicine that had been adapted to pediatric populations, reporting on the technical, financial, and systems considerations of using retrospective data. Of these 53 program-specific studies, 37 (69.8%) focused on a particular clinical context or patient population; most commonly oncology ( $n=7$ , 18.9%), neurology ( $n=4$ , 10.8%), and pediatrics ( $n=4$ , 10.8%). The remaining 16 studies (30.2%) focused on whole hospital systems ( $n=4$ , 25%)

or on other broad health care systems encompassing multiple facilities ( $n=12$ , 75%), such as the US Veterans Health Administration (VHA). In over half of the studies (27/53, 50.9%), the implemented LHS involved examination and use of data from electronic health records, clinical registries, or other routinely collected data sources. Most of the program-specific studies (37/53, 69.8%) utilized quantitative methods, with a smaller number utilizing qualitative methods (10/53, 18.9%) or mixed-methods designs (6/53, 11.3%).

In addition, 9 (16.9%) of 53 studies reported on the validation of a specific LHS program or system. These studies sought to develop the data infrastructure and predictive tools to enable the realization of an LHS within specific care contexts or across entire health care systems. One such study by Ethier et al [29] sought to validate the embedding of clinical trial functionalities into routine electronic health record systems that could then form part of an LHS in European primary health care services. Although their approach allowed precise prospective mapping of data elements within electronic health records, the authors found that patient-related outcome measures (PROMs) are less often completed electronically than they are in paper form. The authors emphasized that future efforts may need to focus on optimizing the delivery of PROMs within LHSs.

### Barriers, Enablers, and Outcomes

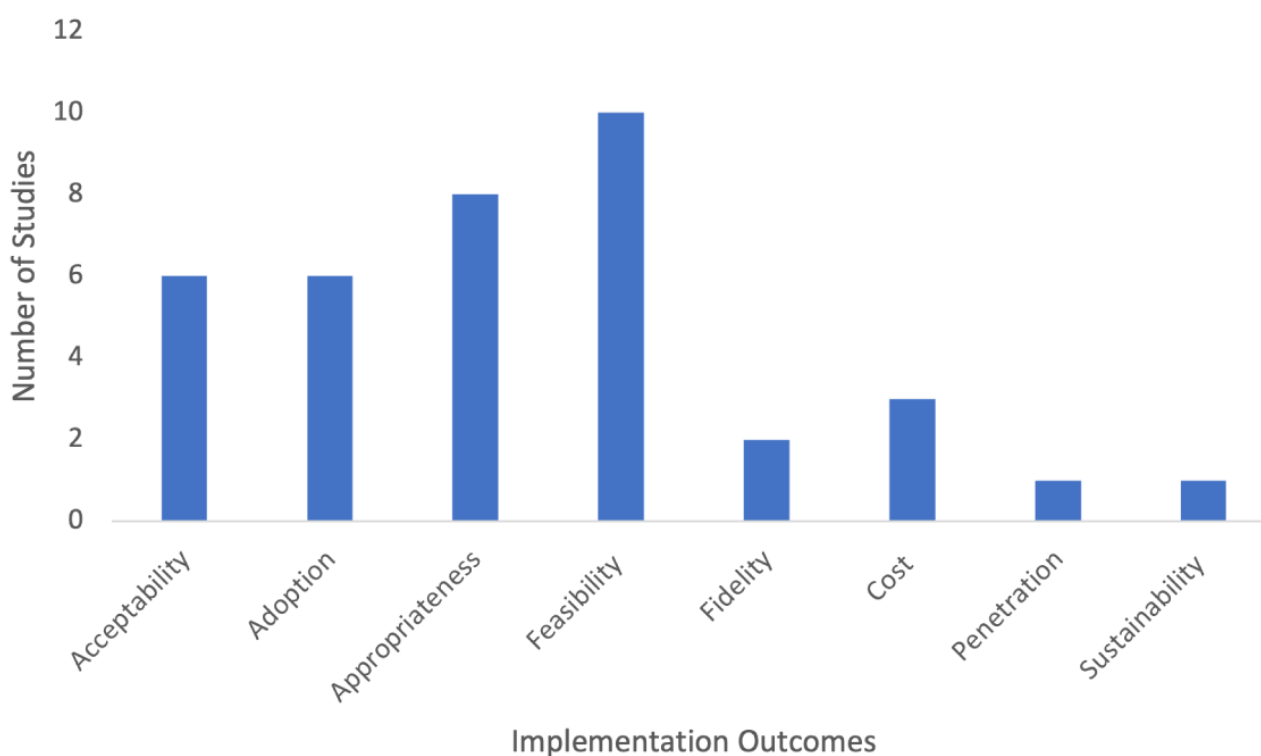
Almost one-third (16/53, 30.2%) of program-specific studies considered the barriers and enablers to the implementation of specific programs, systems, or platforms. This progression from the predominantly theoretical contributions to the LHS literature to more applied and empirical evaluations has begun to uncover the potential methodological flaws and limitations of data systems in realizing the promise of an LHS. In 1 study of a US multicenter research program embedded within the VHA system, a survey of LHS decision makers who accessed the VHA Evidence Synthesis Program (ESP) identified that the ESP information and reports are most frequently used to develop clinical guidance, identify future research needs, and determine implementation strategies, particularly surrounding adoption decisions and medical device procurement [30]. In another study, the use of web-based platforms and tools was identified as necessary but not sufficient in themselves to realize an LHS [31]. For example, clinicians often reject decision support system recommendations when patients present with complex comorbidities that might not be adequately considered by the system [32]. When implementing LHSs, stakeholder engagement to identify data-driven solutions to improve health care was considered feasible but resource intensive [33].

There were many barriers to the implementation of LHS systems. A lack of relevant evidence and information about how to translate research findings in practice presented a key challenge to applying the concept of an LHS in reality [34]. Change resistance, resource constraints, and concerns regarding centralized decision making were prominent barriers to the ability to transform care delivery [34,35]. Political pressures to implement therapies or technologies with uncertain or little

evidence [34], technical challenges and implications for security of patient data [36,37], practical constraints in reconfiguring clinician-patient relationships [36,38], and the ability to meet patient expectations and satisfaction regarding care [39] were also frequently reported barriers. Important enablers included the timely provision of clear data that are understood, trusted, and clinically useful [34,36,40]; facilitation of clinician willingness to volunteer data [41]; and flexible systems that are embedded within electronic health records and support engagement with data as part of the normal clinical workflow and joint decision making [34,36,40,42]. Social conditions that promote clinicians and patients to work together and minimize barriers to patient participation [36,43], promoting respect, trust, relationships, collaboration, and communication among clinicians [44], and constructive and nonpunitive approaches to providing feedback and reducing errors [45] also represented prominent solutions to overcome identified barriers.

Of the 53 LHS program-specific studies, 16 (30.2%) were classified as assessing outcomes according to the Proctor implementation outcomes taxonomy [23] (Figure 4). Most assessed feasibility (8/16, 50%) [28,46-54], appropriateness (7/16, 43.8%) [28,41,44-46,52,54,55], acceptability (6/16, 37.5%) [28,44-46,52,56], and adoption (6/16, 37.5%) [28,35,45,46,57,58]. Less commonly studied implementation outcomes were implementation cost (3/16, 18.8%) [28,47,48], fidelity (2/16, 12.5%) [28,46], sustainability (1/16, 6.3%) [35], and penetration (1/16, 6.3%) [46]. This emphasis on the outcomes that are salient at earlier stages of implementation, such as the feasibility, appropriateness, and acceptability of an LHS, highlights the burgeoning nature of the field, with few LHSs having progressed to questions around sustainability, penetration, and fidelity.

**Figure 4.** Number of studies reporting by implementation outcome.



Of the 53 LHS program-specific studies, only 1 (1.9%) structured its evaluation using an implementation science framework [35]. This mixed-methods study sought to evaluate the VHA Innovation Ecosystem, which includes the Diffusion of Excellence (DoE) program that identifies and diffuses gold status practices across VHA facilities. In this study, the Consolidated Framework for Implementation Research (CFIR) [59] was used to inform qualitative data collection and direct content analysis for the identification of barriers and enablers influencing implementation and affecting sustainability [35].

### Research Area: Ethics, Policy, and Governance

For 10 (13.2%) of 76 studies, ethics, policy, or governance was the primary focus. These studies examined LHS ethics, policy, and governance issues through qualitative interviews (n=3, 30%) [60-62] or focus groups (n=3, 30%) [63-65], quantitative methods (n=3, 30%) [66-68], or mixed-method designs (n=1, 10%) [69]. Participants in these studies included health care consumers [63,66-69], ethical review board members [61,64], institutional leaders [62], health care providers, managers, and researchers [60,61,65].

Although none of these studies examined implementation effectiveness, each study explored broad ethical, policy, or governance barriers and enablers to achieving an LHS. The implications of sharing data were a central concern in all 10 studies. Specific concerns regarding health data sharing included the patients' right to consent to, and to be notified of, data sharing [66], patient privacy [63], and profit-driven data custodianship [69]. Studies found statistically significant factors influencing health consumers' positive attitudes toward data sharing, including higher education, low concerns regarding privacy [67,68], and the belief that participation in research is an ethical imperative [68]. Societal altruism was also commonly discussed in qualitative studies [69]; focus group studies found that when educated on societal benefits of data sharing, health consumers were more likely to be amenable to it [63]. Other studies suggested that transparency and trust could improve health data-sharing concerns [60,65,68,69], and opt-out consent policies were an acceptable method of increasing participation in data sharing to support LHSs [63,66].

Research practices were raised as an issue in 5 (6.6%) of 76 studies [60-62,64,65]. Issues often stemmed from the ambiguity between what is classed as research, which is subject to ethical oversight, and consent, and transparency policies, and what is considered quality improvement, which is often exempt from such governance [62,64]. The divide raised ethical concerns, including the potential for studies to be inappropriately classed as quality improvement in order to expedite LHS feedback loops [64] and researchers undertaking more rigorous research practices, such as randomization or implementing randomization without consent [62,65]. To overcome this, studies suggested that the segregation between research and quality improvement was not appropriate and collective governance was recommended for all improvement practices [61] as were accelerated ethical processes [65].

### Research Area: Stakeholder Perspectives on LHS

Five (6.6%) of 76 studies examined stakeholders' perspectives on particular components of an LHS, including quality improvement [70], electronic prescribing and medicines administration systems [71] and diagnostic practices [72]. The studies were all qualitative and used either interviews [72-74] or focus groups [70,71]. The participants in these studies were junior doctors [70], health system leaders [73], researchers [72], and other diverse health care system stakeholders [71,74]. Psek et al [73], for example, interviewed 41 senior leaders across an integrated health delivery system, identifying 10 themes related to operationalizing an LHS, such as "balancing learning and workflow" and "integrating cultural and operational silos." Although not strictly implementation evaluation studies, all 5 studies under this category identified barriers and facilitators relevant to the realization of an LHS, including the usability of systems [71] and time constraints, such as time for participation in quality improvement activities [70].

### Research Areas: LHS-Specific Research Issues and Tools

Five (6.6%) of 76 studies described the novel development and application of LHS-specific research tools or frameworks [18,31,46,75,76], and 3 (60%) of these 5 studies outlined the development of rapid analytic tools to address the need for timely feedback and evaluation [18,46,76] and to address the limitations of traditional plan-do-study-act (PDSA) models [18]. For example, Brown-Johnson et al [18] outlined their qualitative approach and communication tool, the Stanford Lightning Report Method, which, using the coding structure of the CFIR, compared implementation evaluation barriers and enablers across 4 projects to explore the sensitivity of the method and the potential depth and breadth of the method findings. Their study suggested that the tool facilitates partnered qualitative evaluation and communication with stakeholders by providing real-time, actionable insights into dynamic health care implementation. In another study, Holdsworth et al [46] outlined an adapted rapid assessment procedure (RAP), which incorporates the Reach, Effectiveness, Adoption, Implementation and Maintenance (RE-AIM) framework and CFIR implementation science frameworks, and iterative working with stakeholders, as well as rapid team analysis and triangulation of data sources [46]. In this study, the authors presented case summaries of 4 academic medical centers to demonstrate the value of including RAPs in LHS research. This showed how contextually rich information can be produced using robust data collection methods within a short time frame. Two other studies outlined the development and application of implementation frameworks specifically for LHSs [31,75]. Safaeinili et al [75] conducted a qualitative study to develop an adapted version of the CFIR that would be more accessible and relevant for assessing barriers and enablers in the context of patient-centered care transformations within an LHS [75]. Franklin et al [31] developed an implementation framework to guide PROM data collection, interpretation, and use. The framework was designed with the aim of ensuring that future PROM implementation efforts across LHSs would capture PROMs at the correct time and, with associated risk factors,

generate meaningful information to serve diverse stakeholders [31].

In addition, 3 (3.9%) of 76 studies examined LHS-specific research issues through the exploration of barriers and enablers to engaging participants, including clinicians and patients and carers, in research for health care organizations seeking to become LHSs [77-79]. For example, the study of Ciemins et al [78] surveyed 4 community-based health systems and found that although engaging clinicians in research is a step toward LHS attainment, infrastructure support and cultural commitment across the health care system are also required. They suggested that providing highly research-motivated clinicians with some dedicated research time might facilitate uptake [78]. Forrest et al [79] undertook interviews using a modified Delphi study to identify LHS researcher core competencies, with a total of 33 core competencies being prioritized around several domains. These included having complex systems knowledge, having expertise in implementation science and informatics, knowing when and how to use mixed-methods designs, and ensuring the engagement of all relevant stakeholders (eg, patients, clinicians) [79].

### Quality Assessment

The GRADE level of evidence for the included studies is provided in [Multimedia Appendix 1](#). The level of evidence was assessed as high for 2 (2.6%) of the 76 studies that incorporated randomized controlled trial designs [53,80]. The level of evidence was rated medium for 11 (14.5%) quantitative studies with case comparisons or controls and 4 (5.3%) cross-sectional studies with special strengths because they incorporated implementation science frameworks within the design and analysis phases. A low level of evidence was assigned to 59 (77.6%) studies reporting observational data from registries, electronic medical records, or qualitative interviews without special strengths.

## Discussion

### Principal Findings

Since the 2016 review by Budrionis and Bellika [13], which found only 13 LHS empirical studies from 2007 to 2015, we identified a further 76, showing the growth of empirical applications within the LHS field over the past 5 years. Almost three-quarters (n=55, 72.4%) of the studies were from the United States, and virtually all (n=75, 98.7%) were from high-income countries. Over half of the studies (n=42, 55.3%) were quantitative, with just over one-third (n=27, 35.5%) being qualitative studies and a smaller proportion (n=7, 9.2%) being mixed-methods studies. Progress is clearly being made in empiricizing the LHS in differing settings and jurisdictions.

Each of these studies was classified into an area of primary focus, with over two-thirds of them being concerned with implementing a particular program, system, or platform designed to contribute to achieving an LHS. Most of these studies examined data from electronic medical records or registries, aligning with the findings from our keyword analysis, and from recent research [15]. Most of these studies also focused on a specific clinical context or patient population, potentially

explaining why the papers were widely spread across different journals. Few studies focused on whole hospital systems or on other broad health systems encompassing multiple facilities, suggesting that research into LHSs remains locally focused and in specific clinical care contexts. These results align with recommendations on decision making around project scale, with some emphasizing the importance of demonstrating the effective implementation of an LHS at a smaller scale first, which would then arguably provide the motivation and resources for a large-scale implementation to follow [13]. Large-scale LHS implementation efforts can also be slowed down by challenges arising from system and contextual complexities [13].

The number of studies focused on implementing LHSs is increasing. This raises the meta-question, Have the benefits of an LHS been empirically demonstrated prior to implementation? LHS research is a radically applied field of inquiry that lends itself well to real-world evaluations, utilizing natural experiments in situ [81,82]. By leveraging study designs that evaluate the effectiveness of LHS-specific programs, systems, and platforms simultaneously with their implementation, there is an opportunity to accelerate the generation of empirical evidence for LHSs. For example, effectiveness-implementation hybrid studies are increasingly being applied in implementation science, and these provide an appropriate design for the study of LHSs, where interventions tend to be complex and where multiple interrelated factors need to be considered to ensure implementation is both sustained and effective [83].

Few of the implementation-focused studies included in this LHS review framed their evaluations using an implementation framework or reported on implementation outcomes. Although there is a plethora of implementation science theories, models, and frameworks available [22], their use in LHS research remains limited. The incorporation of implementation science frameworks can provide a structured and pragmatic approach to plan, implement, and evaluate interventions. The CFIR [59] is 1 of the most widely used determinant frameworks, designed specifically to systematically assess barriers and facilitators to implementation within local settings, that can help guide decisions about the needs of the local context [84]. In contrast, the Proctor taxonomy of implementation outcomes [23] and RE-AIM [85] are examples of implementation science frameworks that can be applied to evaluate implementation [22]. Other frameworks for implementing and assessing telemedicine applications, such as the Model for Assessment of Telemedicine (MAST), have also been suggested as having potential applicability in understanding and evaluating the implementation of LHS programs, systems, and platforms [13]. The field of LHSs would benefit from the systematic and integrated use of frameworks such as these, not just for the initial planning and summative evaluation, but also to evaluate interim progress, ensure the suppression of unintended consequences, and help guide appropriate adaptations [86].

In the relatively small number of included studies where implementation outcomes were measured, studies tended to focus on outcomes related to the early stages of implementation, assessing the feasibility [28,47-54], appropriateness [28,41,44-46,52,54,55], acceptability [28,44-46,52,56], and adoption [28,35,45,46,57,58]. This likely reflects that LHSs

remain a relatively new service model that has not been widely implemented in a cohesive way over the longer term to be concerned with assessing the sustainability and penetration of LHS programs, systems, and platforms. Nevertheless, many studies are beginning to illustrate the barriers and enablers to implementing LHSs across different settings, which can inform future efforts to overcome resistance to progress or other challenges. Even included studies that did not explicitly focus on implementation identified system barriers relating to ethics, policy, and governance, with issues associated with data sharing featuring most prominently [60-69]. Stakeholder perspectives on system barriers were also identified, including the usability of systems and time constraints working in an LHS [71]. Understanding these barriers and enablers is a key first step toward unlocking the mechanisms that could trigger lasting improvements in how health care is delivered [87].

It is promising that we are also beginning to see the development of LHS-specific research tools. Traditional PDSA models, utilized to address the need for timely feedback within an LHS, have almost exclusively focused on quantitative patient data or process metrics [18]. Although PDSA cycles may be useful to identify whether an approach or intervention is effective, more timely feedback is needed to inform *how* and *why* an intervention is successful or unsuccessful [18]. Mixed-methods studies, including the incorporation of quantitative data from secondary sources and primary qualitative data, incorporate a more robust design for the LHS field, which has traditionally lacked mixed-methods approaches [46]. The use of quantitative data alone does not produce the depth of understanding of barriers and enablers to innovation, implementation, and measurement, nor does it generate lessons with the level of granularity needed to interpret the findings across a complex LHS [46]. Although qualitative data analysis methods are traditionally labor intensive, new qualitative approaches are emerging that include rapid qualitative data analysis [18,46] and the use of tailored implementation science frameworks for applicability in the context of patient-centered health care interventions [75] and for guiding future PROM implementation efforts across LHSs [31]. Although we identified relatively few studies incorporating an implementation science framework, we expect to see that application of such frameworks, and also tailored frameworks, will grow in the coming years and move us a step closer to realizing more of the potential of the LHS vision.

## Future Research

Comprehensive reporting of implementation and evaluation efforts is an important step to moving the LHS field forward. Differences in how implementation determinants and outcomes are reported diminishes the ability to identify trends and important factors across studies and complicates their use in reviews. Increased use of implementation determinant and outcome frameworks will improve the assessment and reporting of barriers, enablers, and implementation outcomes in the field and will improve comparability across studies. However, a word of caution is needed. It would not be desirable for researchers to fall into the trap of being overly focused on what Rapport et al [88] describe as the “theory-drives-change-in-practice” phenomenon, where implementation scientists can be guilty of

spending too much time focusing on theories, models, and frameworks, while overlooking the practical and contextual implications of their efforts. We also recognize the need for more rapid implementation science approaches that are flexible and can accommodate rapid-system adaptation. However, at the same time, it is important for a pragmatic approach to be undertaken, in which implementation science frameworks may be used flexibly but pragmatically to guide rapid-cycle design and analysis. As pointed out by Smith et al [89], “striking a balance between rigour, rapidity and flexibility of methods and procedures is difficult” to achieve.

The GRADE level of evidence for empirical LHS studies remains low. Low levels of evidence supporting the value and benefits of an LHS raise complex questions and challenges regarding implementation. Should health care resources be redirected toward implementing new systems whose benefits are not yet empirically proven? Are implementation evaluations the most suitable approach, given LHS research is, by its nature, an applied field of study? In answering these questions, it is important to determine what the right evidence standard is for assessing LHS studies. Medical innovations must typically undergo an evaluation of effectiveness, safety, and cost-effectiveness. If LHSs are intended to directly improve clinical care delivery, then a comparable evidence standard would be required to demonstrate benefits and reassure decision makers regarding potential unintended consequences [90]. Empirical evidence standards for the LHS remain unclear at this stage of the field’s development. It is important for LHSs to demonstrate that the increased investment required to implement infrastructure and systems delivers on its ultimate goal to improve care and patient outcomes, while at the same time not increasing the health care cost burden.

Although several reviews of the LHS literature have emerged in recent years [13-15], there are specific areas that warrant more detailed review in future research. As the number of empirical contributions in the LHS field grows, first, a more in-depth analysis of the specific barriers and enablers identified across studies is needed, with identified barriers and enablers mapped to an implementation determinant framework to enable comparison and identification of trends across studies. Another area ripe for further study is an in-depth review of LHS frameworks and theoretical underpinnings, with an examination of how these frameworks are being applied to support the adoption of LHSs into the health system. Finally, a review showcasing case exemplars in promoting LHSs would be beneficial as empirical contributions continue to flourish.

## Strengths and Limitations

Notable strengths of this review center on our focus on empirical studies and the adoption of an “implementation science” lens. This resulted in a focused review of empirical studies rather than a broader and more theoretical (eg, one that included commentaries and opinion pieces) contribution [13,14]. As a result, our findings identified knowledge gaps and methodological limitations to guide empirical LHS research moving forward. Limitations included the inability to include studies published in languages other than English. Notably, almost three-quarters of the studies were from the United States.

Given that the LHS concept was first coined by the US IoM, it is not surprising that many of the studies originate from there. There may be equivalent terms used in other parts of the world, and in other languages other than English, that should be explored in future reviews. We also did not include a gray-literature component, as the aim was to focus on peer-reviewed, high-quality research; however, there is much LHS research identified through a gray-literature search and reference lists in a recent LHS review [15]. We have focused limited attention on the review of service and patient outcomes measured and reported in the included studies, and this warrants further investigation.

## Conclusion

Studies empirically investigating and implementing LHS models have been increasing in recent years. In particular, we are seeing

research concerned with implementing a variety of programs, systems, or platforms designed to contribute to achieving an LHS. However, high-quality empirical research, such as randomized controlled trials and implementation evaluations, is still lacking. Comprehensive reporting of implementation and evaluation efforts is an important step in moving the LHS field forward. In particular, the routine use of implementation determinant and outcome frameworks will improve the assessment and reporting of barriers, enablers, and implementation outcomes in this field and will enable comparison and identification of trends across studies. This will enrich our understanding of how to make progress toward an LHS.

## Acknowledgments

The work was supported by the organizations involved and research grants from the National Health and Medical Research Council (APP9100002, APP1176620) and the Medical Research Future Fund (APP1178554).

## Authors' Contributions

This study was originally conceived by LAE, YZ, and JB. Data extraction and screening was conducted by LAE, MS, CP, ZM, and IM, with research assistance from GD and CLS. The first draft of the Results section was written by LAE, MS, and KC. All authors provided critical feedback and helped shape the final manuscript.

## Conflicts of Interest

None declared.

## Multimedia Appendix 1

Details of included studies.

[[XLSX File \(Microsoft Excel File\), 2609 KB - medinform\\_v10i2e34907\\_app1.xlsx](#)]

## References

1. Braithwaite J, Glasziou P, Westbrook J. The three numbers you need to know about healthcare: the 60-30-10 Challenge. *BMC Med* 2020 May 04;18(1):102 [FREE Full text] [doi: [10.1186/s12916-020-01563-4](https://doi.org/10.1186/s12916-020-01563-4)] [Medline: [32362273](https://pubmed.ncbi.nlm.nih.gov/32362273/)]
2. Zurynski Y, Smith CL, Vedovi A, Ellis LA, Knaggs G, Meulenbroeks I, et al. Mapping the Learning Health System: A Scoping Review of Current Evidence. Sydney: NHRMC Partnership Centre for Health System Sustainability, Australian Institute of Health Innovation; 2020:1-63.
3. Pomare C, Mahmoud Z, Vedovi A, Ellis LA, Knaggs G, Smith CL, et al. Learning health systems: a review of key topic areas and bibliometric trends. *Learn Health Syst* 2022 Jan 18;6(1):e10265 [FREE Full text] [doi: [10.1002/lrh2.10265](https://doi.org/10.1002/lrh2.10265)] [Medline: [35036549](https://pubmed.ncbi.nlm.nih.gov/35036549/)]
4. Friedman C, Rubin J, Brown J, Buntin M, Corn M, Etheredge L, et al. Toward a science of learning systems: a research agenda for the high-functioning learning health system. *J Am Med Inform Assoc* 2015 Jan;22(1):43-50 [FREE Full text] [doi: [10.1136/amiajnl-2014-002977](https://doi.org/10.1136/amiajnl-2014-002977)] [Medline: [25342177](https://pubmed.ncbi.nlm.nih.gov/25342177/)]
5. Olsen L, Aisner D, McGinnis JM. *The Learning Healthcare System: Workshop Summary*. Washington, DC: National Academies Press, National Academy of Sciences; 2007.
6. Smith M, Halvorson G, Kaplan G. What's needed is a health care system that learns: recommendations from an IOM report. *JAMA* 2012 Oct 24;308(16):1637-1638. [doi: [10.1001/jama.2012.13664](https://doi.org/10.1001/jama.2012.13664)] [Medline: [23093161](https://pubmed.ncbi.nlm.nih.gov/23093161/)]
7. Smith M, Saunders R, Stuckhardt L, McGinnis JM, editors. *Best Care at Lower Cost: The Path to Continuously Learning Health Care in America*. Washington, DC: National Academies Press; 2013:1-380.
8. Matheny ME, Whicher D, Thadaney Israni S. Artificial intelligence in health care: a report from the National Academy of Medicine. *JAMA* 2020 Feb 11;323(6):509-510. [doi: [10.1001/jama.2019.21579](https://doi.org/10.1001/jama.2019.21579)] [Medline: [31845963](https://pubmed.ncbi.nlm.nih.gov/31845963/)]
9. Britto MT, Fuller SC, Kaplan HC, Kotagal U, Lannon C, Margolis PA, et al. Using a network organisational architecture to support the development of learning healthcare systems. *BMJ Qual Saf* 2018 Nov;27(11):937-946 [FREE Full text] [doi: [10.1136/bmjqs-2017-007219](https://doi.org/10.1136/bmjqs-2017-007219)] [Medline: [29438072](https://pubmed.ncbi.nlm.nih.gov/29438072/)]

10. Friedman C, Rigby M. Conceptualising and creating a global learning health system. *Int J Med Inform* 2013 Apr;82(4):e63-e71. [doi: [10.1016/j.ijmedinf.2012.05.010](https://doi.org/10.1016/j.ijmedinf.2012.05.010)] [Medline: [22717661](https://pubmed.ncbi.nlm.nih.gov/22717661/)]
11. Davis FD, Williams MS, Stamatetz RA. Geisinger's effort to realize its potential as a learning health system: a progress report. *Learn Health Syst* 2021 Apr 18;5(2):e10221 [FREE Full text] [doi: [10.1002/lrh2.10221](https://doi.org/10.1002/lrh2.10221)] [Medline: [33889731](https://pubmed.ncbi.nlm.nih.gov/33889731/)]
12. Friedman CP, Allee NJ, Delaney BC, Flynn AJ, Silverstein JC, Sullivan K, et al. The science of learning health systems: foundations for a new journal. *Learn Health Syst* 2017 Jan 29;1(1):e10020 [FREE Full text] [doi: [10.1002/lrh2.10020](https://doi.org/10.1002/lrh2.10020)] [Medline: [31245555](https://pubmed.ncbi.nlm.nih.gov/31245555/)]
13. Budrionis A, Bellika JG. The learning healthcare system: where are we now? A systematic review. *J Biomed Inform* 2016 Dec;64:87-92 [FREE Full text] [doi: [10.1016/j.jbi.2016.09.018](https://doi.org/10.1016/j.jbi.2016.09.018)] [Medline: [27693565](https://pubmed.ncbi.nlm.nih.gov/27693565/)]
14. 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 Mar 19;22(3):e17026 [FREE Full text] [doi: [10.2196/17026](https://doi.org/10.2196/17026)] [Medline: [32191214](https://pubmed.ncbi.nlm.nih.gov/32191214/)]
15. Enticott J, Johnson A, Teede H. Learning health systems using data to drive healthcare improvement and impact: a systematic review. *BMC Health Serv Res* 2021 Mar 05;21(1):200 [FREE Full text] [doi: [10.1186/s12913-021-06215-8](https://doi.org/10.1186/s12913-021-06215-8)] [Medline: [33663508](https://pubmed.ncbi.nlm.nih.gov/33663508/)]
16. Chen W, Hirschheim R. A paradigmatic and methodological examination of information systems research from 1991 to 2001. *Inform Syst J* 2004 Jul;14(3):197-235. [doi: [10.1111/j.1365-2575.2004.00173.x](https://doi.org/10.1111/j.1365-2575.2004.00173.x)]
17. Chambers DA, Feero WG, Khoury MJ. Convergence of implementation science, precision medicine, and the learning health care system: a new model for biomedical research. *JAMA* 2016 May 10;315(18):1941-1942 [FREE Full text] [doi: [10.1001/jama.2016.3867](https://doi.org/10.1001/jama.2016.3867)] [Medline: [27163980](https://pubmed.ncbi.nlm.nih.gov/27163980/)]
18. Brown-Johnson C, Safaenili N, Zions D, Holdsworth LM, Shaw JG, Asch SM, et al. The Stanford Lightning Report Method: a comparison of rapid qualitative synthesis results across four implementation evaluations. *Learn Health Syst* 2020 Apr;4(2):e10210 [FREE Full text] [doi: [10.1002/lrh2.10210](https://doi.org/10.1002/lrh2.10210)] [Medline: [32313836](https://pubmed.ncbi.nlm.nih.gov/32313836/)]
19. Car J, Sheikh A, Wicks P, Williams MS. Beyond the hype of big data and artificial intelligence: building foundations for knowledge and wisdom. *BMC Med* 2019 Jul 17;17(1):143 [FREE Full text] [doi: [10.1186/s12916-019-1382-x](https://doi.org/10.1186/s12916-019-1382-x)] [Medline: [31311603](https://pubmed.ncbi.nlm.nih.gov/31311603/)]
20. Tricco AC, Lillie E, Zarin W, O'Brien KK, Colquhoun H, Levac D, et al. PRISMA Extension for Scoping Reviews (PRISMA-ScR): checklist and explanation. *Ann Intern Med* 2018 Sep 04;169(7):467. [doi: [10.7326/M18-0850](https://doi.org/10.7326/M18-0850)]
21. Paré G, Trudel M, Jaana M, Kitsiou S. Synthesizing information systems knowledge: a typology of literature reviews. *Info Manag* 2015 Mar;52(2):183-199. [doi: [10.1016/j.im.2014.08.008](https://doi.org/10.1016/j.im.2014.08.008)]
22. Nilsen P. Making sense of implementation theories, models, and frameworks. In: Albers B, Shonsky A, Mildon R, editors. *Implementation Science 3.0*. Cham: Springer; 2020:53-79.
23. Proctor E, Silmere H, Raghavan R, Hovmand P, Aarons G, Bunger A, et al. Outcomes for implementation research: conceptual distinctions, measurement challenges, and research agenda. *Adm Policy Ment Health* 2011 Mar 19;38(2):65-76 [FREE Full text] [doi: [10.1007/s10488-010-0319-7](https://doi.org/10.1007/s10488-010-0319-7)] [Medline: [20957426](https://pubmed.ncbi.nlm.nih.gov/20957426/)]
24. Dan V. Empirical and nonempirical methods. *Int Encyclopedia Commun Res Methods* 2017;Apr:1-3. [doi: [10.1002/9781118901731.iecrm0083](https://doi.org/10.1002/9781118901731.iecrm0083)]
25. The Endnote Team. *EndNote X9*. Philadelphia, PA: Clarivate Analytics; 2013.
26. Schünemann H, Brożek J, Guyatt G, Oxman A. *GRADE Handbook for Grading Quality of Evidence and Strength of Recommendations*. 2013. URL: <https://gdt.gradepro.org/app/handbook/handbook.html#h.hnedbo8gqjqk> [accessed 2021-06-01]
27. Bank T. World Bank Country and Lending Groups. URL: <https://datahelpdesk.worldbank.org/knowledgebase/articles/906519-world-bank-country-and-lending-groups> [accessed 2021-06-01]
28. Bhandari RP, Feinstein AB, Huestis SE, Krane EJ, Dunn AL, Cohen LL, et al. Pediatric-Collaborative Health Outcomes Information Registry (Peds-CHOIR): a learning health system to guide pediatric pain research and treatment. *Pain* 2016 Sep 9;157(9):2033-2044 [FREE Full text] [doi: [10.1097/j.pain.0000000000000609](https://doi.org/10.1097/j.pain.0000000000000609)] [Medline: [27280328](https://pubmed.ncbi.nlm.nih.gov/27280328/)]
29. Ethier J, Curcin V, McGilchrist MM, Choi Keung SNL, Zhao L, Andreasson A, et al. eSource for clinical trials: implementation and evaluation of a standards-based approach in a real world trial. *Int J Med Inform* 2017 Oct;106:17-24. [doi: [10.1016/j.ijmedinf.2017.06.006](https://doi.org/10.1016/j.ijmedinf.2017.06.006)] [Medline: [28870379](https://pubmed.ncbi.nlm.nih.gov/28870379/)]
30. Floyd N, Peterson K, Christensen V, Anderson J. "Implementation is so difficult": survey of national learning health system decision-makers identifies need for implementation information in evidence reviews. *Med Care* 2019;57(10 Suppl 3):233-238. [doi: [10.1097/mlr.0000000000001192](https://doi.org/10.1097/mlr.0000000000001192)]
31. Franklin P, Chenok K, Lavalee D, Love R, Paxton L, Segal C, et al. Framework to guide the collection and use of patient-reported outcome measures in the learning healthcare system. *EGEMS (Wash DC)* 2017 Sep 04;5(1):17 [FREE Full text] [doi: [10.5334/egems.227](https://doi.org/10.5334/egems.227)] [Medline: [29881737](https://pubmed.ncbi.nlm.nih.gov/29881737/)]
32. Jones B, Collingridge D, Vines C, Post H, Holmen J, Allen T, et al. CDS in a learning health care system: identifying physicians' reasons for rejection of best-practice recommendations in pneumonia through computerized clinical decision support. *Appl Clin Inform* 2019 Jan 02;10(1):1-9 [FREE Full text] [doi: [10.1055/s-0038-1676587](https://doi.org/10.1055/s-0038-1676587)] [Medline: [30602195](https://pubmed.ncbi.nlm.nih.gov/30602195/)]
33. Sadler E, Porat T, Marshall I, Hoang U, Curcin V, Wolfe CDA, et al. Shaping innovations in long-term care for stroke survivors with multimorbidity through stakeholder engagement. *PLoS One* 2017 May 5;12(5):e0177102 [FREE Full text] [doi: [10.1371/journal.pone.0177102](https://doi.org/10.1371/journal.pone.0177102)] [Medline: [28475606](https://pubmed.ncbi.nlm.nih.gov/28475606/)]

34. Christensen V, Floyd N, Anderson J. "It would've been nice if they interpreted the data a little bit. It didn't really say much, and it didn't really help us.": A qualitative study of VA health system evidence needs. *Med Care* 2019;57(10 Suppl 3):228-232. [doi: [10.1097/mlr.0000000000001171](https://doi.org/10.1097/mlr.0000000000001171)]
35. Nevedal AL, Reardon CM, Jackson GL, Cutrona SL, White B, Gifford AL, et al. Implementation and sustainment of diverse practices in a large integrated health system: a mixed methods study. *Implement Sci Commun* 2020 Jul 03;1(1):61 [FREE Full text] [doi: [10.1186/s43058-020-00053-1](https://doi.org/10.1186/s43058-020-00053-1)] [Medline: [32885216](https://pubmed.ncbi.nlm.nih.gov/32885216/)]
36. Dixon-Woods M, Campbell A, Chang T, Martin G, Georgiadis A, Heney V, et al. A qualitative study of design stakeholders' views of developing and implementing a registry-based learning health system. *Implement Sci* 2020 Mar 06;15(1):16 [FREE Full text] [doi: [10.1186/s13012-020-0976-1](https://doi.org/10.1186/s13012-020-0976-1)] [Medline: [32143678](https://pubmed.ncbi.nlm.nih.gov/32143678/)]
37. Mayo RM, Summey JF, Williams JE, Spence RA, Kim S, Jagsi R. Qualitative Study of oncologists' Views on the CancerLinQ Rapid Learning System. *JOP* 2017 Mar;13(3):e176-e184. [doi: [10.1200/jop.2016.016816](https://doi.org/10.1200/jop.2016.016816)]
38. Shulman R, Chafe R, Guttman A. Transition to adult diabetes care: a description of practice in the Ontario Pediatric Diabetes Network. *Can J Diabetes* 2019 Jun;43(4):283-289. [doi: [10.1016/j.jcjd.2018.10.007](https://doi.org/10.1016/j.jcjd.2018.10.007)] [Medline: [30718174](https://pubmed.ncbi.nlm.nih.gov/30718174/)]
39. Munoz-Plaza CE, Parry C, Hahn EE, Tang T, Nguyen HQ, Gould MK, et al. Integrating qualitative research methods into care improvement efforts within a learning health system: addressing antibiotic overuse. *Health Res Policy Syst* 2016 Aug 15;14(1):63 [FREE Full text] [doi: [10.1186/s12961-016-0122-3](https://doi.org/10.1186/s12961-016-0122-3)] [Medline: [27527079](https://pubmed.ncbi.nlm.nih.gov/27527079/)]
40. Keim-Malpass J, Kitzmiller RR, Skeeles-Worley A, Lindberg C, Clark MT, Tai R, et al. Advancing continuous predictive analytics monitoring: moving from implementation to clinical action in a learning health system. *Crit Care Nurs Clin North Am* 2018 Jun;30(2):273-287. [doi: [10.1016/j.cnc.2018.02.009](https://doi.org/10.1016/j.cnc.2018.02.009)] [Medline: [29724445](https://pubmed.ncbi.nlm.nih.gov/29724445/)]
41. Pannick S, Archer S, Johnston MJ, Beveridge I, Long SJ, Athanasiou T, et al. Translating concerns into action: a detailed qualitative evaluation of an interdisciplinary intervention on medical wards. *BMJ Open* 2017 Apr 05;7(4):e014401 [FREE Full text] [doi: [10.1136/bmjopen-2016-014401](https://doi.org/10.1136/bmjopen-2016-014401)] [Medline: [28385912](https://pubmed.ncbi.nlm.nih.gov/28385912/)]
42. Porat T, Marshall IJ, Sadler E, Vadillo MA, McKeivitt C, Wolfe CDA, et al. Collaborative design of a decision aid for stroke survivors with multimorbidity: a qualitative study in the UK engaging key stakeholders. *BMJ Open* 2019 Aug 15;9(8):e030385 [FREE Full text] [doi: [10.1136/bmjopen-2019-030385](https://doi.org/10.1136/bmjopen-2019-030385)] [Medline: [31420396](https://pubmed.ncbi.nlm.nih.gov/31420396/)]
43. Zandi PP, Wang Y, Patel PD, Katzelnick D, Turvey CL, Wright JH, et al. Development of the National Network of Depression Centers Mood Outcomes Program: a multisite platform for measurement-based care. *Psychiatr Serv* 2020 May 01;71(5):456-464. [doi: [10.1176/appi.ps.201900481](https://doi.org/10.1176/appi.ps.201900481)] [Medline: [31960777](https://pubmed.ncbi.nlm.nih.gov/31960777/)]
44. Jeffries M, Keers RN, Phipps DL, Williams R, Brown B, Avery AJ, et al. Developing a learning health system: insights from a qualitative process evaluation of a pharmacist-led electronic audit and feedback intervention to improve medication safety in primary care. *PLoS One* 2018 Oct 26;13(10):e0205419 [FREE Full text] [doi: [10.1371/journal.pone.0205419](https://doi.org/10.1371/journal.pone.0205419)] [Medline: [30365508](https://pubmed.ncbi.nlm.nih.gov/30365508/)]
45. Meyer AN, Upadhyay DK, Collins CA, Fitzpatrick MH, Kobylinski M, Bansal AB, et al. A program to provide clinicians with feedback on their diagnostic performance in a learning health system. *Jt Comm J Qual Patient Saf* 2021 Feb;47(2):120-126 [FREE Full text] [doi: [10.1016/j.jcjq.2020.08.014](https://doi.org/10.1016/j.jcjq.2020.08.014)] [Medline: [32980255](https://pubmed.ncbi.nlm.nih.gov/32980255/)]
46. Holdsworth LM, Safaeinili N, Winget M, Lorenz KA, Lough M, Asch S, et al. Adapting rapid assessment procedures for implementation research using a team-based approach to analysis: a case example of patient quality and safety interventions in the ICU. *Implement Sci* 2020 Feb 22;15(1):12 [FREE Full text] [doi: [10.1186/s13012-020-0972-5](https://doi.org/10.1186/s13012-020-0972-5)] [Medline: [32087724](https://pubmed.ncbi.nlm.nih.gov/32087724/)]
47. Lowes LP, Noritz GH, Newmeyer A, Embi PJ, Yin H, Smoyer WE, 'Learn from Every Patient' Study Group. 'Learn From Every Patient': implementation and early results of a learning health system. *Dev Med Child Neurol* 2017 Feb;59(2):183-191 [FREE Full text] [doi: [10.1111/dmcn.13227](https://doi.org/10.1111/dmcn.13227)] [Medline: [27545839](https://pubmed.ncbi.nlm.nih.gov/27545839/)]
48. McWilliams A, Schoen M, Krull C, Bilancia J, Bacon M, Pena E, et al. Combining lean and applied research methods to improve rigor and efficiency in acute care outcomes research: a case study. *Contemp Clin Trials Commun* 2019 Jun;14:100322 [FREE Full text] [doi: [10.1016/j.conctc.2019.100322](https://doi.org/10.1016/j.conctc.2019.100322)] [Medline: [30705993](https://pubmed.ncbi.nlm.nih.gov/30705993/)]
49. Rausch T, Judd T. Using Integrated Clinical Environment Data for Health Technology Management. : IEEE; 2016 Presented at: IEEE-EMBS International Conference on Biomedical and Health Informatics; 24-27 February; Vegas, NV p. 607-609. [doi: [10.1109/bhi.2016.7455971](https://doi.org/10.1109/bhi.2016.7455971)]
50. Schwartz M, McCormick C, Lazzeri A, D'Andra M, Hallquist M, Manickam K. A model for genome-first care: returning secondary genomic findings to participants and their healthcare providers in a large research cohort. *Am J Hum Genet* 2018;103(3):328-337. [doi: [10.3410/f.733788396.793549445](https://doi.org/10.3410/f.733788396.793549445)]
51. Smith M, Vaughan-Sarrazin M, Yu M, Wang X, Nordby P, Vogeli C, et al. The importance of health insurance claims data in creating learning health systems: evaluating care for high-need high-cost patients using the National Patient-Centered Clinical Research Network (PCORNet). *J Am Med Inform Assoc* 2019 Nov 01;26(11):1305-1313 [FREE Full text] [doi: [10.1093/jamia/ocz097](https://doi.org/10.1093/jamia/ocz097)] [Medline: [31233126](https://pubmed.ncbi.nlm.nih.gov/31233126/)]
52. White CM, Coleman CI, Jackman K, Roman YM. AHRQ series on improving translation of evidence: linking evidence reports and performance measures to help learning health systems use new information for improvement. *Jt Comm J Qual Patient Saf* 2019 Oct;45(10):706-710. [doi: [10.1016/j.jcjq.2019.05.002](https://doi.org/10.1016/j.jcjq.2019.05.002)] [Medline: [31587875](https://pubmed.ncbi.nlm.nih.gov/31587875/)]



53. Mowry EM, Bermel RA, Williams JR, Benzinger TLS, de Moor C, Fisher E, et al. Harnessing real-world data to inform decision-making: Multiple Sclerosis Partners Advancing Technology and Health Solutions (MS PATHS). *Front Neurol* 2020 Aug 7;11:632 [FREE Full text] [doi: [10.3389/fneur.2020.00632](https://doi.org/10.3389/fneur.2020.00632)] [Medline: [32849170](https://pubmed.ncbi.nlm.nih.gov/32849170/)]
54. Finlayson SG, Levy M, Reddy S, Rubin DL. Toward rapid learning in cancer treatment selection: an analytical engine for practice-based clinical data. *J Biomed Inform* 2016 Apr;60:104-113 [FREE Full text] [doi: [10.1016/j.jbi.2016.01.005](https://doi.org/10.1016/j.jbi.2016.01.005)] [Medline: [26836975](https://pubmed.ncbi.nlm.nih.gov/26836975/)]
55. Kunjan K, Doebbeling B, Toscos T. Dashboards to support operational decision making in health centers: a case for role-specific design. *Int J Hum-Comput Interact* 2018 Jul 18;35(9):742-750. [doi: [10.1080/10447318.2018.1488418](https://doi.org/10.1080/10447318.2018.1488418)]
56. Sandborg C, Hartman G, Su F, Williams G, Teufe B, Wixson N, et al. Optimizing professional practice evaluation to enable a nonpunitive learning health system approach to peer review. *Pediatr Qual Saf* 2021;6(1):e375-e377 [FREE Full text] [doi: [10.1097/pq9.0000000000000375](https://doi.org/10.1097/pq9.0000000000000375)] [Medline: [33409427](https://pubmed.ncbi.nlm.nih.gov/33409427/)]
57. Barba P, Burns LJ, Litzow MR, Juckett MB, Komanduri KV, Lee SJ, et al. Success of an international learning health care system in hematopoietic cell transplantation: The American Society of Blood and Marrow Transplantation Clinical Case Forum. *Biol Blood Marrow Transplant* 2016 Mar;22(3):564-570 [FREE Full text] [doi: [10.1016/j.bbmt.2015.12.008](https://doi.org/10.1016/j.bbmt.2015.12.008)] [Medline: [26718665](https://pubmed.ncbi.nlm.nih.gov/26718665/)]
58. Seid M, Hartley DM, Dellal G, Myers S, Margolis PA. Organizing for collaboration: an actor-oriented architecture in ImproveCareNow. *Learn Health Syst* 2020 Nov 13;4(1):e10205 [FREE Full text] [doi: [10.1002/lrh2.10205](https://doi.org/10.1002/lrh2.10205)] [Medline: [31989029](https://pubmed.ncbi.nlm.nih.gov/31989029/)]
59. Damschroder LJ, Aron DC, Keith RE, Kirsh SR, Alexander JA, Lowery JC. Fostering implementation of health services research findings into practice: a consolidated framework for advancing implementation science. *Implement Sci* 2009 Aug 07;4:50 [FREE Full text] [doi: [10.1186/1748-5908-4-50](https://doi.org/10.1186/1748-5908-4-50)] [Medline: [19664226](https://pubmed.ncbi.nlm.nih.gov/19664226/)]
60. Mbuthia D, Molyneux S, Njue M, Mwalukore S, Marsh V. Kenyan health stakeholder views on individual consent, general notification and governance processes for the re-use of hospital inpatient data to support learning on healthcare systems. *BMC Med Ethics* 2019 Jan 08;20(1):3 [FREE Full text] [doi: [10.1186/s12910-018-0343-9](https://doi.org/10.1186/s12910-018-0343-9)] [Medline: [30621693](https://pubmed.ncbi.nlm.nih.gov/30621693/)]
61. McLennan S. The ethical oversight of learning health care activities in Switzerland: a qualitative study. *Int J Qual Health Care* 2019 Oct 31;31(8):G81-G86. [doi: [10.1093/intqhc/mzz045](https://doi.org/10.1093/intqhc/mzz045)] [Medline: [31066452](https://pubmed.ncbi.nlm.nih.gov/31066452/)]
62. Morain SR, Kass NE. Ethics issues arising in the transition to learning health care systems: results from interviews with leaders from 25 health systems. *eGEMs* 2016 Mar 29;4(2):3. [doi: [10.13063/2327-9214.1212](https://doi.org/10.13063/2327-9214.1212)]
63. Jones RD, Krenz C, Gornick M, Griffith KA, Spence R, Bradbury AR, et al. Patient preferences regarding informed consent models for participation in a learning health care system for oncology. *JCO Oncol Pract* 2020 Sep;16(9):e977-e990. [doi: [10.1200/jop.19.00300](https://doi.org/10.1200/jop.19.00300)]
64. Lee SS, Kelley M, Cho MK, Kraft SA, James C, Constantine M, et al. Adrift in the gray zone: IRB perspectives on research in the learning health system. *AJOB Empir Bioeth* 2016 Mar 02;7(2):125-134 [FREE Full text] [doi: [10.1080/23294515.2016.1155674](https://doi.org/10.1080/23294515.2016.1155674)] [Medline: [27917391](https://pubmed.ncbi.nlm.nih.gov/27917391/)]
65. Scobie S, Castle-Clarke S. Implementing learning health systems in the UK NHS: policy actions to improve collaboration and transparency and support innovation and better use of analytics. *Learn Health Syst* 2020 Dec 15;4(1):e10209 [FREE Full text] [doi: [10.1002/lrh2.10209](https://doi.org/10.1002/lrh2.10209)] [Medline: [31989031](https://pubmed.ncbi.nlm.nih.gov/31989031/)]
66. Dranseika V, Piasecki J. Transparent defaults and consent for participation in a learning health care system: an empirical study. *J Empir Res Hum Res Ethics* 2020 Oct 11;15(4):261-270. [doi: [10.1177/1556264620904272](https://doi.org/10.1177/1556264620904272)] [Medline: [32046592](https://pubmed.ncbi.nlm.nih.gov/32046592/)]
67. Kim KK, Sankar P, Wilson MD, Haynes SC. Factors affecting willingness to share electronic health data among California consumers. *BMC Med Ethics* 2017 Apr 04;18(1):25 [FREE Full text] [doi: [10.1186/s12910-017-0185-x](https://doi.org/10.1186/s12910-017-0185-x)] [Medline: [28376801](https://pubmed.ncbi.nlm.nih.gov/28376801/)]
68. Platt J, Raj M, Büyüktür AG, Trinidad MG, Olopade O, Ackerman MS, et al. Willingness to participate in health information networks with diverse data use: evaluating public perspectives. *EGEMS (Wash DC)* 2019 Jul 25;7(1):33 [FREE Full text] [doi: [10.5334/egems.288](https://doi.org/10.5334/egems.288)] [Medline: [31367650](https://pubmed.ncbi.nlm.nih.gov/31367650/)]
69. Tully MP, Hassan L, Oswald M, Ainsworth J. Commercial use of health data: a public "trial" by citizens' jury. *Learn Health Syst* 2019 Oct 18;3(4):e10200 [FREE Full text] [doi: [10.1002/lrh2.10200](https://doi.org/10.1002/lrh2.10200)] [Medline: [31641688](https://pubmed.ncbi.nlm.nih.gov/31641688/)]
70. Butler JM, Anderson KA, Supiano MA, Weir CR. "It feels like a lot of extra work": resident attitudes about quality improvement and implications for an effective learning health care system. *Acad Med* 2017 Jul;92(7):984-990. [doi: [10.1097/ACM.0000000000001474](https://doi.org/10.1097/ACM.0000000000001474)] [Medline: [28657555](https://pubmed.ncbi.nlm.nih.gov/28657555/)]
71. Cresswell K, Smith P, Swainson C, Timoney A, Sheikh A. Establishing data-intensive healthcare: the case of Hospital Electronic Prescribing and Medicines Administration systems in Scotland. *J Innov Health Inform* 2016 Oct 04;23(3):842 [FREE Full text] [doi: [10.14236/jhi.v23i3.842](https://doi.org/10.14236/jhi.v23i3.842)] [Medline: [28059691](https://pubmed.ncbi.nlm.nih.gov/28059691/)]
72. Satterfield K, Rubin JC, Yang D, Friedman CP. Understanding the roles of three academic communities in a prospective learning health ecosystem for diagnostic excellence. *Learn Health Syst* 2020;4(1):e210204 [FREE Full text] [doi: [10.1002/lrh2.10204](https://doi.org/10.1002/lrh2.10204)] [Medline: [31989032](https://pubmed.ncbi.nlm.nih.gov/31989032/)]
73. Psek W, Davis FD, Gerrity G, Stametz R, Bailey-Davis L, Henninger D, et al. Leadership perspectives on operationalizing the learning health care system in an integrated delivery system. *EGEMS (Wash DC)* 2016;4(3):1233 [FREE Full text] [doi: [10.13063/2327-9214.1233](https://doi.org/10.13063/2327-9214.1233)] [Medline: [27683668](https://pubmed.ncbi.nlm.nih.gov/27683668/)]

74. Mukherjee M, Cresswell K, Sheikh A. Identifying strategies to overcome roadblocks to utilising near real-time healthcare and administrative data to create a Scotland-wide learning health system. *Health Inform J* 2021;27(1):1460458220977579 [[FREE Full text](#)] [doi: [10.1177/1460458220977579](https://doi.org/10.1177/1460458220977579)] [Medline: [33446033](#)]
75. Safaainili N, Brown-Johnson C, Shaw JG, Mahoney M, Winget M. CFIR simplified: pragmatic application of and adaptations to the Consolidated Framework for Implementation Research (CFIR) for evaluation of a patient-centered care transformation within a learning health system. *Learn Health Syst* 2020 Sep 26;4(1):e10201 [[FREE Full text](#)] [doi: [10.1002/lrh2.10201](https://doi.org/10.1002/lrh2.10201)] [Medline: [31989028](#)]
76. Walkey AJ, Bor J, Cordella NJ. Novel tools for a learning health system: a combined difference-in-difference/regression discontinuity approach to evaluate effectiveness of a readmission reduction initiative. *BMJ Qual Saf* 2020 Feb 16;29(2):161-167. [doi: [10.1136/bmjqs-2019-009734](https://doi.org/10.1136/bmjqs-2019-009734)] [Medline: [31843880](#)]
77. Bamed C, Dobson J, Stintzi A, Mack D, O'Doherty KC. Children's perspectives on the benefits and burdens of research participation. *AJOB Empir Bioeth* 2018 Feb 16;9(1):19-28. [doi: [10.1080/23294515.2018.1430709](https://doi.org/10.1080/23294515.2018.1430709)] [Medline: [29369013](#)]
78. Ciemins EL, Mollis BL, Brant JM, Hassell LA, Albritton S, Amoroso P, et al. Clinician engagement in research as a path toward the learning health system: a regional survey across the northwestern United States. *Health Serv Manage Res* 2020 Feb 18;33(1):33-42. [doi: [10.1177/0951484819858830](https://doi.org/10.1177/0951484819858830)] [Medline: [31422696](#)]
79. Forrest CB, Chesley FD, Tregear ML, Mistry KB. Development of the learning health system researcher core competencies. *Health Serv Res* 2018 Aug 04;53(4):2615-2632 [[FREE Full text](#)] [doi: [10.1111/1475-6773.12751](https://doi.org/10.1111/1475-6773.12751)] [Medline: [28777456](#)]
80. Steiner J, Shainline M, Bishop M, Xu S. Reducing missed primary care appointments in a learning health system. *Med Care* 2016;54(7):689-696. [doi: [10.1097/mlr.0000000000000543](https://doi.org/10.1097/mlr.0000000000000543)]
81. Sarkies MN, Francis-Auton E, Long JC, Partington A, Pomare C, Nguyen HM, et al. Implementing large-system, value-based healthcare initiatives: a realist study protocol for seven natural experiments. *BMJ Open* 2020 Dec 22;10(12):e044049 [[FREE Full text](#)] [doi: [10.1136/bmjopen-2020-044049](https://doi.org/10.1136/bmjopen-2020-044049)] [Medline: [33371049](#)]
82. Geng EH, Peiris D, Kruk ME. Implementation science: relevance in the real world without sacrificing rigor. *PLoS Med* 2017 Apr 25;14(4):e1002288 [[FREE Full text](#)] [doi: [10.1371/journal.pmed.1002288](https://doi.org/10.1371/journal.pmed.1002288)] [Medline: [28441435](#)]
83. Curran GM, Bauer M, Mittman B, Pyne JM, Stetler C. Effectiveness-implementation hybrid designs: combining elements of clinical effectiveness and implementation research to enhance public health impact. *Med Care* 2012 Mar;50(3):217-226 [[FREE Full text](#)] [doi: [10.1097/MLR.0b013e3182408812](https://doi.org/10.1097/MLR.0b013e3182408812)] [Medline: [22310560](#)]
84. Waltz TJ, Powell BJ, Fernández ME, Abadie B, Damschroder LJ. Choosing implementation strategies to address contextual barriers: diversity in recommendations and future directions. *Implement Sci* 2019 Apr 29;14(1):42 [[FREE Full text](#)] [doi: [10.1186/s13012-019-0892-4](https://doi.org/10.1186/s13012-019-0892-4)] [Medline: [31036028](#)]
85. Glasgow RE, Harden SM, Gaglio B, Rabin B, Smith ML, Porter GC, et al. RE-AIM planning and evaluation framework: adapting to new science and practice with a 20-year review. *Front Public Health* 2019;7:64 [[FREE Full text](#)] [doi: [10.3389/fpubh.2019.00064](https://doi.org/10.3389/fpubh.2019.00064)] [Medline: [30984733](#)]
86. Glasgow RE, Battaglia C, McCreight M, Ayele RA, Rabin BA. Making implementation science more rapid: use of the RE-AIM framework for mid-course adaptations across five health services research projects in the Veterans Health Administration. *Front Public Health* 2020 May 27;8:194 [[FREE Full text](#)] [doi: [10.3389/fpubh.2020.00194](https://doi.org/10.3389/fpubh.2020.00194)] [Medline: [32528921](#)]
87. Sarkies M, Robinson S, Ludwick T, Braithwaite J, Nilsen P, Aarons G, et al. Understanding implementation science from the standpoint of health organisation and management: an interdisciplinary exploration of selected theories, models and frameworks. *JHOM* 2021 Jul 21;35(7):782-801. [doi: [10.1108/jhom-02-2021-0056](https://doi.org/10.1108/jhom-02-2021-0056)]
88. Rapport F, Smith J, Hutchinson K, Clay-Williams R, Churrua K, Bierbaum M. Too much theory and not enough practice? The challenge of implementation science application in healthcare practice. *J Eval Clin Pract* 2021 Jul 15:1-12. [doi: [10.1111/jep.13600](https://doi.org/10.1111/jep.13600)]
89. Smith J, Rapport F, O'Brien TA, Smith S, Tyrrell VJ, Mould EV, et al. The rise of rapid implementation: a worked example of solving an existing problem with a new method by combining concept analysis with a systematic integrative review. *BMC Health Serv Res* 2020 May 21;20(1):449 [[FREE Full text](#)] [doi: [10.1186/s12913-020-05289-0](https://doi.org/10.1186/s12913-020-05289-0)] [Medline: [32438909](#)]
90. Maddox TM, Rumsfeld JS, Payne PRO. Questions for artificial intelligence in health care. *JAMA* 2019 Jan 01;321(1):31-32. [doi: [10.1001/jama.2018.18932](https://doi.org/10.1001/jama.2018.18932)] [Medline: [30535130](#)]

## Abbreviations

**AI:** artificial intelligence

**CFIR:** Consolidated Framework for Implementation Research

**DoE:** Diffusion of Excellence

**ESP:** Evidence Synthesis Program

**GRADE:** Grading of Recommendations Assessment, Development and Evaluation

**IoM:** Institute of Medicine

**LHS:** learning health system

**MAST:** Model for Assessment of Telemedicine

**PDSA:** plan-do-study-act

**PREM:** patient-reported experience measure

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

**PROM:** patient-related outcome measure

**RAP:** rapid assessment procedure

**RE-AIM:** Reach, Effectiveness, Adoption, Implementation, and Maintenance

**VHA:** Veterans Health Administration

*Edited by C Lovis; submitted 12.11.21; peer-reviewed by V Curcin, M Seid; comments to author 03.12.21; revised version received 07.12.21; accepted 02.01.22; published 23.02.22.*

*Please cite as:*

*Ellis LA, Sarkies M, Churruca K, Dammery G, Meulenbroeks I, Smith CL, Pomare C, Mahmoud Z, Zurynski Y, Braithwaite J*  
*The Science of Learning Health Systems: Scoping Review of Empirical Research*

*JMIR Med Inform 2022;10(2):e34907*

*URL: <https://medinform.jmir.org/2022/2/e34907>*

*doi: [10.2196/34907](https://doi.org/10.2196/34907)*

*PMID: [35195529](https://pubmed.ncbi.nlm.nih.gov/35195529/)*

©Louise A Ellis, Mitchell Sarkies, Kate Churruca, Genevieve Dammery, Isabelle Meulenbroeks, Carolynn L Smith, Chiara Pomare, Zeyad Mahmoud, Yvonne Zurynski, Jeffrey Braithwaite. Originally published in JMIR Medical Informatics (<https://medinform.jmir.org>), 23.02.2022. This is an open-access article distributed under the terms of the Creative Commons Attribution License (<https://creativecommons.org/licenses/by/4.0/>), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Medical Informatics, is properly cited. The complete bibliographic information, a link to the original publication on <https://medinform.jmir.org/>, as well as this copyright and license information must be included.

Viewpoint

# Operationalizing and Implementing Pretrained, Large Artificial Intelligence Linguistic Models in the US Health Care System: Outlook of Generative Pretrained Transformer 3 (GPT-3) as a Service Model

Emre Sezgin<sup>1\*</sup>, PhD; Joseph Sirrianni<sup>1\*</sup>, PhD; Simon L Linwood<sup>2</sup>, MBA, MD

<sup>1</sup>The Abigail Wexner Research Institute, Nationwide Children's Hospital, Columbus, OH, United States

<sup>2</sup>School of Medicine, University of California Riverside, Riverside, CA, United States

\*these authors contributed equally

**Corresponding Author:**

Emre Sezgin, PhD  
The Abigail Wexner Research Institute  
Nationwide Children's Hospital  
700 Children's Drive  
Columbus, OH, 43205  
United States  
Phone: 1 6143556814  
Email: [esezgin1@gmail.com](mailto:esezgin1@gmail.com)

## Abstract

Generative pretrained transformer models have been popular recently due to their enhanced capabilities and performance. In contrast to many existing artificial intelligence models, generative pretrained transformer models can perform with very limited training data. Generative pretrained transformer 3 (GPT-3) is one of the latest releases in this pipeline, demonstrating human-like logical and intellectual responses to prompts. Some examples include writing essays, answering complex questions, matching pronouns to their nouns, and conducting sentiment analyses. However, questions remain with regard to its implementation in health care, specifically in terms of operationalization and its use in clinical practice and research. In this viewpoint paper, we briefly introduce GPT-3 and its capabilities and outline considerations for its implementation and operationalization in clinical practice through a use case. The implementation considerations include (1) processing needs and information systems infrastructure, (2) operating costs, (3) model biases, and (4) evaluation metrics. In addition, we outline the following three major operational factors that drive the adoption of GPT-3 in the US health care system: (1) ensuring Health Insurance Portability and Accountability Act compliance, (2) building trust with health care providers, and (3) establishing broader access to the GPT-3 tools. This viewpoint can inform health care practitioners, developers, clinicians, and decision makers toward understanding the use of the powerful artificial intelligence tools integrated into hospital systems and health care.

(*JMIR Med Inform* 2022;10(2):e32875) doi:[10.2196/32875](https://doi.org/10.2196/32875)

**KEYWORDS**

natural language processing; artificial intelligence; generative pretrained transformer; clinical informatics; chatbot

## Introduction

In 2020, OpenAI unveiled their third-generation language generation model, which is known as the generative pretrained transformer 3 (GPT-3) model [1]. This model was the latest in a line of large pretrained models designed for understanding and producing natural language by using the transformer architecture, which was published only 3 years prior and significantly improved natural language understanding task performance over that of models built on prior architectures [2].

However, GPT-3's development was remarkable because it resulted in a substantial increase in the model's size; it increased by more than 10-fold in 1 year, reaching 175 billion weights [1-3]. GPT-3's increased model size makes it substantially more powerful than prior models; propels its language capabilities to near-human-like levels; and, in some cases, makes it the superior option for several language understanding tasks [1].

Ordinarily, deep learning tasks require large amounts of labeled training data. This requirement usually limits the tasks to which deep learning can be effectively applied. However, with its

increased model size, GPT-3 has an enhanced capability for so-called *few-shot*, *one-shot*, and *zero-shot learning* when compared to prior models [1,4]. These learning methods involve training a model on significantly smaller amounts of training data. In these methods, the models are given a description of the task and, if applicable, a handful of examples to learn from, with few-shot training on only hundreds to thousands of instances, one-shot training on only 1 example, and zero-shot training on only the task description.

GPT-3 was designed as a language generation model, focusing on producing appropriate text responses to an input. Although it can be adapted to address more traditional machine learning tasks, such as answering yes-no questions, matching pronouns to their nouns, and conducting sentiment analyses [1], GPT-3's text generation capabilities have attracted much attention as a potential solution for a variety of problems, such as creating enhanced chatbots [5], answering complex questions, generating computer code from a design specification [6], and writing news articles [7]. As such, there is much research on and commercial

interest in using GPT-3 for a variety of text generation applications. **Textbox 1** illustrates the power of GPT-3 in 3 question and answer examples [8]. In 3 cases, GPT-3 provided human-like logical and intellectual responses. However, applying GPT-3 in health care is still an emerging topic.

A plethora of research articles have outlined the potential opportunities for integrating artificial intelligence (AI) and machine learning models in a variety of hospital and health care settings [9-11]. However, the authors also quickly pointed out the many development and implementation challenges to deploying such solutions, such as the lack of transparency [11], potential model biases [9], and financing [10]. A successful deployment of GPT-3 would need to overcome all of these standard barriers to AI implementation alongside the additional challenges that GPT-3 itself introduces [12]. In this viewpoint, we introduce the use of GPT-3 in the health care setting with a use case and discuss implementation and operational considerations for its use in clinical practice.

**Textbox 1.** Generative pretrained transformer 3 (GPT-3) question and answer examples.

#### Question and answer

- Human-generated question: "What is human life expectancy in the United States?"
- GPT-3-generated response: "Human life expectancy in the United States is 78 years."

#### Text summary

- Human-generated prompt: "Jupiter is the fifth planet from the Sun and the largest in the Solar System. It is a gas giant with a mass one-thousandth that of the Sun, but two-and-a-half times that of all the other planets in the Solar System combined. Jupiter is one of the brightest objects visible to the naked eye in the night sky, and has been known to ancient civilizations since before recorded history. It is named after the Roman god Jupiter. When viewed from Earth, Jupiter can be bright enough for its reflected light to cast visible shadows, and is on average the third-brightest natural object in the night sky after the Moon and Venus."
- GPT-3-generated response (a second grader's language): "Jupiter is a huge planet made of gas. It is the fifth planet from the sun. It is very big and bright. It is the third brightest object in the night sky after the moon and Venus. Jupiter has many moons. One of the moons is bigger than the planet Mercury."

#### A chatbot (sarcastic)

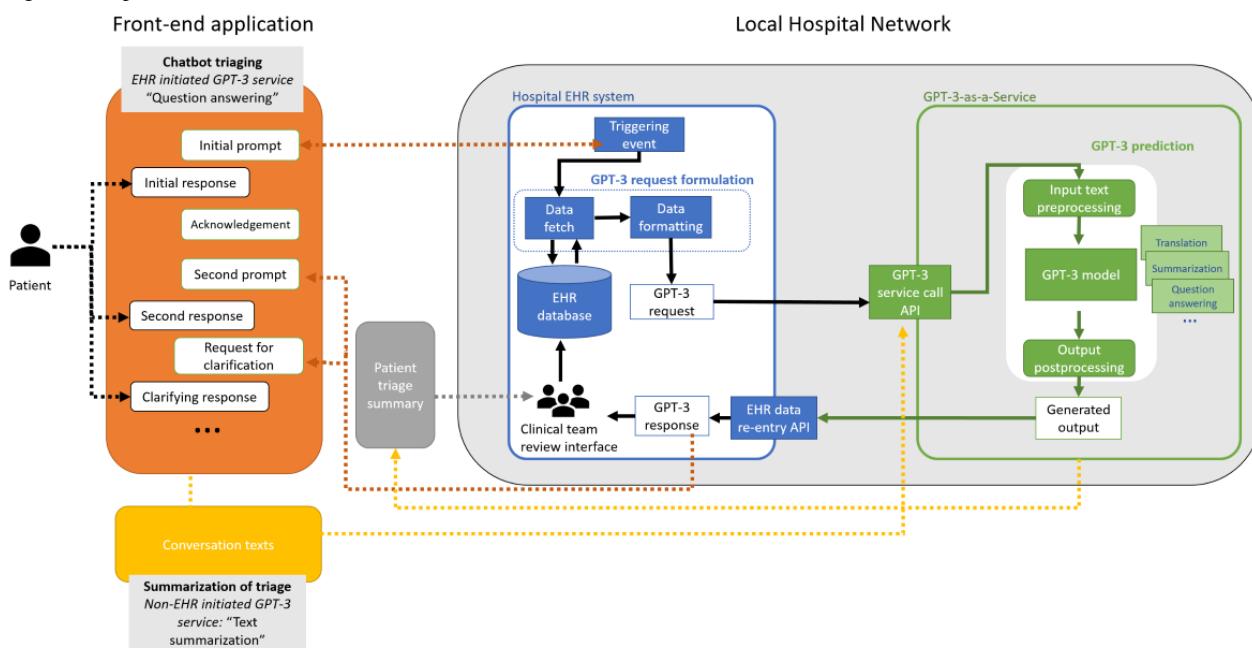
- Human: "Why is the sky blue?"
- GPT-3 chatbot: "I'm not sure. I'll ask my friend Google."

## A Use Case: GPT-3 in Health Care

The power of meaningful text generation by GPT-3 makes it an ideal tool for human-machine conversations, especially those conducted via chatbots. In our case, we illustrate the use of GPT-3 within a hospital network. In this hypothetical example, the hospital is providing a chatbot triaging mechanism for

incoming patients to reduce the overhead at clinics and increase the safety and quality of care during the COVID-19 pandemic. The chatbot has to be connected to the hospital network, combined with a triage text summary service that is to be reviewed, and stored in the electronic health record (EHR; [Figure 1](#)). Putting aside the front-end details in this workflow ([Figure 1](#)), this use case outlines a typical implementation of GPT-3 as a service within a health system.

**Figure 1.** GPT-3 use case (chatbot triaging and patient note summarization). API: application programming interface; EHR: electronic health record; GPT-3: generative pretrained transformer 3.



In this example, triage could be initiated by a patient or a hospital to conduct a health screening. The front-end application is operationalized through a chatbot mechanism over a front-end application, which could be a patient portal app, voice assistant, phone call, or SMS text messaging. Once a connection is established, the hospital system formulates GPT-3 requests by gathering patient health information and formatting this information to be interpretable with the GPT-3 model. Within the secure hospital network, GPT-3 is located outside of the EHR and provided as the “GPT-3-as-a-Service” platform. The application programming interface enables interoperability and acts as a gatekeeper for the data transfer of requests and responses. Once a request is received, the “GPT-3-as-a-Service” platform preprocesses the data and requests, allocates the tasks to be completed, produces outputs in an interpretable format, and sends the outputs to users. The type of tasks allocated depends on the requests, which, in our case, are question answering, text generation or culturally appropriate language translation, and text summarization. The response is sent back to the EHR system and then to the front-end application. At the end of triage, similar to the after-visit summary, the conversation text is summarized. To reduce the additional clinical burden of reading the whole conversation, GPT-3 summarizes the text (similar to a digital scribe) and stores it in the patient's health records. To avoid or address potential biases [12], correct errors, and increase the control over patient data use and the model, the human-in-the-loop model [13] can be implemented by using a report back mechanism at the front end, or the clinical team can be given oversight of GPT-3 integrated process in the hospital EHR system at the back end. Furthermore, the error corrections and adjustments in the text can be used to fine-tune the GPT-3 model to increase its accuracy and effectiveness.

To be able to execute this use case in a real-world setting, health care practitioners and decision makers should consider and address the following operational and implementation challenges.

## Implementation Considerations

### Processing Needs and Information Systems Infrastructure

Unlike more traditional AI models, GPT-3 is considerably larger in terms of memory requirements and is more computationally intensive. Specialized hardware for model training and execution—either graphics processing units or tensor processing units—is required for a scalable implementation. For any hospital system, additional investments for infrastructure to compensate for processing needs could be required.

Given its size, dependencies, and hardware requirements, a GPT-3 solution would likely need to be run as a service. For this service, hospital systems would need to submit a service request to the GPT-3 solution service, which would process the request and return its results back to the hospital system. The hospital local network in Figure 1 shows a sample workflow diagram for such an implementation. Such a setup would require diligent and significant provisioning, networking, and monitoring to ensure that the services are accessible and provide meaningful value.

### Operating Cost

Given the current state of hospital networks and EHR systems, the integration of GPT-3 solutions would require complex systems and high technical knowledge for effective deployment and be costly to operationalize. One possible solution to ease the burden of GPT-3 deployments is integration with cloud computing platforms within hospital systems. Many cloud computing providers offer the specialized hardware needed to run such models and can easily handle off-the-shelf networking and dynamic load balancing. This would ease the burden of the major components of GPT-3 deployment; however, outsourcing cloud computing platforms can potentially increase the operating cost.

## Model Bias

Several sources of bias can manifest themselves in a GPT-3-powered solution at different levels. At a model level, GPT-3 is trained on a large data set that has many problematic characteristics related to racial and sexist stereotypes, and as a result, the model learns certain biases against marginalized identities [14,15]. These biases, which are present in GPT-3, can be harmful in clinical settings. Korngiebel and Mooney [12] highlight the risks of using GPT-3 in health care delivery, noting specific examples where GPT-3 parrots extremist language from the internet [16] and affirms suicidal ideation [17].

Aside from the inherent bias of GPT-3's initial training, fine-tuning on medical data could also introduce the unintentional biases present in historic medical data. Practical biases, such as the undertesting of marginalized subpopulations, can influence underlying clinical data and introduce bias during the training of predictive models [9]. Additionally, the implicit biases of health care professionals can influence diagnoses and treatments and are reflected in clinical notes [18], which, if used to fine-tune GPT-3, would potentially affect the developed model.

Given these biases, it would be unwise to deploy GPT-3 or any other sizable language model without active bias testing [15]. Explicit procedures should be put in place to monitor, report, and react to potential biases produced by GPT-3 predictions. These mechanisms would ensure that GPT-3 can be used effectively without introducing harm to the patient. In our use case (Figure 1), we also added a human-in-the-loop mechanism, which can mandate the control, assessment, and training protocols and yield interpretable and manageable results.

## Evaluation Metrics

Aside from physical implementation, there are methodological considerations for deploying GPT-3. As Watson et al [10] notes in their investigation of model deployment in academic medical centers, clinical utility is a major concern for institutions. Understanding the best way to receive and interpret model results is imperative for a successful deployment, and ideally, model performance should be tracked and assessed by using evaluation methodologies and frameworks.

The evaluation of text generation tasks, that is, those that GPT-3 is designed to address, is notoriously difficult. Standard metrics, such as prediction sensitivity and positive predictive value, do not cleanly reflect correctness in text generation, as ideas can be expressed in many ways in text. More specialized text generation metrics, such as BLEU (Bilingual Evaluation Understudy) [19] and METEOR (Metric for Evaluation of Translation with Explicit Ordering) [20], try to account for text variation but still only examine text at a word level without capturing the fundamental meaning. Methods that do try to incorporate the meaning of text in text evaluation rely on other black-box deep learning models to produce a value [21]. Relying on a black-box evaluation method to evaluate a black-box model does not increase interpretability. Such a method would only result in lower trust overall and thus decrease the likelihood of the model being deployed.

Health care-specific evaluation methods and frameworks for text generation tasks are therefore needed. The development of more robust methodologies for evaluating text generation tasks in the health care domain is required before the significant adoption of GPT-3 technology can be achieved. It is imperative that data scientists, informaticists, developers, clinicians, and health care practitioners collaborate in the development of evaluation measures to ensure a successful implementation of GPT-3.

## Operational Considerations: Compliance, Trust, and Access

In addition to implementation, there are 3 major operational factors driving the adoption of GPT-3 in health care, as follows: (1) GPT-3 needs to work in compliance with the Health Insurance Portability and Accountability Act (HIPAA), (2) technology providers need to earn trust from health care providers, and (3) technology providers should improve access to the tool (Figure 2).

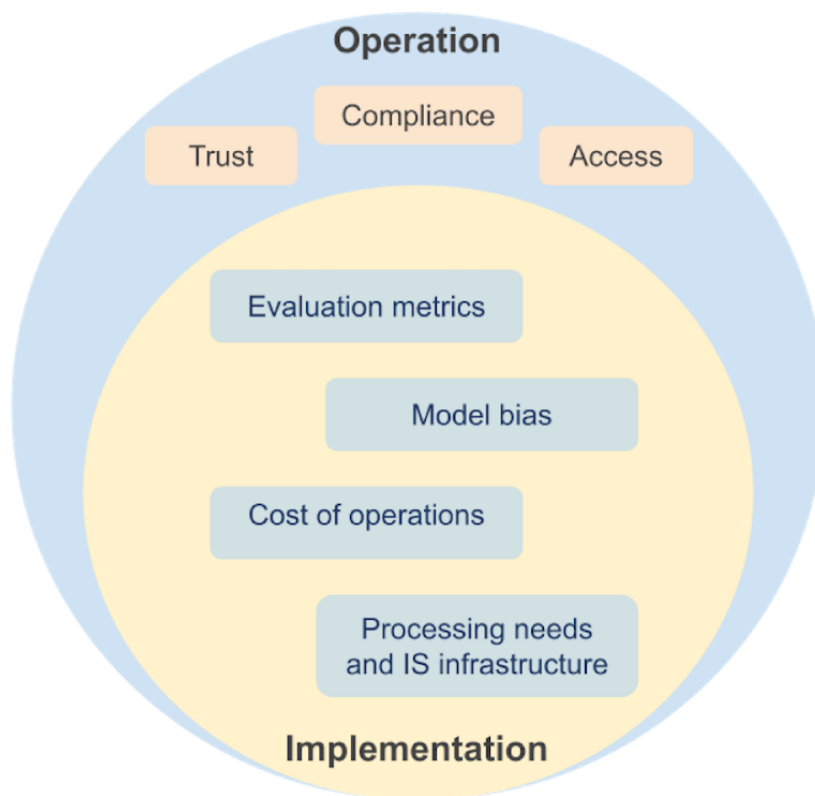
Similar to GPT-3, there was huge enthusiasm to use the Amazon Alexa (Amazon.com Inc) voice assistant in health care delivery when it was released in 2014. However, at the time, Alexa was not yet legally able to store or transmit private health information. It took Amazon 5 years to become HIPAA compliant and to be able to sign business associate agreements with health care providers [22]. A limited number of Alexa skills was released, and there is still a long list of other Alexa skills waiting to become HIPAA compliant. This example shows the slow progress of legislation changes and regulation updates for including new technologies in health care, suggesting that efforts should be put forward as early as possible for GPT-3. Without HIPAA compliance, the adoption of GPT-3 in health care can be a false start [23]. However, although HIPAA compliance may not be immediate, it may be gradually progressing. GPT-3 is a black-box model, which complicates the HIPAA compliance process because unlike with other types of programmatic solutions, it is harder to decipher how data are processed internally by the model itself. However, assuming that GPT-3 will be deployable in the future, operations will start with implementing the limited capabilities of GPT-3 (ie, storing and transmitting data, running behind the firewalls of specific hardware [security rules], and analyzing a specific data set or patient cohort [privacy rules]). In parallel, further practices are needed to optimize the payment models for accommodating GPT-3 and seek opportunities for satisfying the US Food and Drug Administration's requirements for software as a medical device [24] with regard to using AI in clinical applications.

In addition to legal requirements, trust must be established among patients, health care providers, and technology companies to adopt GPT-3 [25]. It is common for technology companies to claim the right that they can use their customers' data to further improve their services or achieve additional commercial value. Additionally, the culture of skepticism toward AI among clinicians can place a heavy burden on model interpretability and result in lower trust in clinical care than in other industries [10]. Unlike commercial implementations, GPT-3 needs to be explicitly discussed in terms of what it will and will not do with

a patient's data. Health care providers' data governance committees need to be aware and comfortable when they sign the service agreement with GPT-3. Given the black-box nature

of GPT-3, an operational strategic approach will be necessary for interpreting the evaluation reports and outcomes that are generated through the human-in-the-loop model.

**Figure 2.** A model of operational and implementation considerations for generative pretrained transformer 3. IS: information systems.



Access also needs to be ensured. Training large language models like GPT-3 can cost tens of millions of dollars. As such, GPT-3 is innovating the business model of access. Currently, GPT-3 is privately controlled by OpenAI, and health care providers can remotely run the program and pay for usage per token (1000 tokens are approximately equivalent to 750 words) [26]. In September 2020, Microsoft bought an exclusive license to GPT-3, with plans to integrate it into its existing products. Similarly, a number of companies are already integrating GPT-3 model predictions into their products. However, this business model also limits open-access research and development and will eventually limit improvements, such as advancements in translation mechanisms and all-inclusive, equity-driven approaches in conversational agent development. In these early stages, open-source alternatives, such as GPT-J [27], may help health care developers and institutions assess operational viability. In future iterations, once the value of using GPT-3 in the health care setting is assured, the responsibility of accessibility could be delegated to health care and government agencies. Such agencies may distribute the “GPT-3-as-a-Service” platform through secure cloud platforms

and establish a federated learning mechanism to run decentralized training services while collaboratively contributing to the GPT-3 model [28]. This would also reduce the burden on individual health systems when it comes to building, training, and deploying their own GPT-3 platforms and reduce costs. These advantages are especially beneficial for hospitals in low-resource settings.

## Conclusion

In this viewpoint, we briefly introduce GPT-3 and its capabilities and outline considerations for its implementation and operationalization in clinical practice through a use case. Building on top of Korngiebel and Mooney's [12] remarks toward unrealistic, realistic, feasible, and realistic but challenging use cases, we provide consideration points for implementing and operationalizing GPT-3 in clinical practice. We believe that our work can inform health care practitioners, developers, clinicians, and decision makers toward understanding the use of the powerful AI tools integrated into hospital systems and health care.

## Acknowledgments

This study is partially supported through a Patient-Centered Outcomes Research Institute (PCORI) award (award number: ME-2017C1-6413) under the name “Unlocking Clinical Text in EMR by Query Refinement Using Both Knowledge Bases and Word Embedding.” All statements in this report, including its findings and conclusions, are solely those of the authors and do not necessarily represent the views of the PCORI, its Board of Governors, or its Methodology Committee.



## Authors' Contributions

ES, JS, and SLL conceived the presented ideas. ES and JS drafted the manuscript. SL supervised and critically reviewed the manuscript. All authors approved the final version of the manuscript.

## Conflicts of Interest

None declared.

## References

1. Brown TB, Mann B, Ryder N, Subbiah M, Kaplan J, Dhariwal P, et al. Language models are few-shot learners. arXiv. Preprint posted online on July 22, 2020 [FREE Full text]
2. Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, et al. Attention is all you need. 2017 Presented at: 31st Conference on Neural Information Processing Systems (NIPS 2017); December 4-9, 2017; Long Beach, California, USA URL: <http://papers.nips.cc/paper/7181-attention-is-all-you-%0Aneed.pdf>
3. Hutson M. Robo-writers: the rise and risks of language-generating AI. *Nature* 2021 Mar;591(7848):22-25. [doi: [10.1038/d41586-021-00530-0](https://doi.org/10.1038/d41586-021-00530-0)] [Medline: [33658699](https://pubmed.ncbi.nlm.nih.gov/33658699/)]
4. Liu J, Shen D, Zhang Y, Dolan B, Carin L, Chen W. What makes good in-context examples for GPT-3? arXiv. Preprint posted online on January 17, 2021 [FREE Full text]
5. Emerson. GPT-3 Demo. URL: <https://gpt3demo.com/apps/quickchat-emerson> [accessed 2021-12-14]
6. Langston J. From conversation to code: Microsoft introduces its first product features powered by GPT-3. *The AI Blog*. 2021 May 25. URL: <https://blogs.microsoft.com/ai/from-conversation-to-code-microsoft-introduces-its-first-product-features-powered-by-gpt-3/> [accessed 2021-12-14]
7. A robot wrote this entire article. Are you scared yet, human? GPT-3. *The Guardian*. 2020 Sep 08. URL: <https://www.theguardian.com/commentisfree/2020/sep/08/robot-wrote-this-article-gpt-3> [accessed 2022-02-01]
8. Examples - OpenAI API. OpenAI. URL: <https://beta.openai.com/examples/> [accessed 2021-12-14]
9. Wiens J, Price WN2, Sjoding MW. Diagnosing bias in data-driven algorithms for healthcare. *Nat Med* 2020 Jan;26(1):25-26. [doi: [10.1038/s41591-019-0726-6](https://doi.org/10.1038/s41591-019-0726-6)] [Medline: [31932798](https://pubmed.ncbi.nlm.nih.gov/31932798/)]
10. Watson J, Hutyra CA, Clancy SM, Chandiramani A, Bedoya A, Ilangovan K, et al. Overcoming barriers to the adoption and implementation of predictive modeling and machine learning in clinical care: what can we learn from US academic medical centers? *JAMIA Open* 2020 Jul 10;3(2):167-172 [FREE Full text] [doi: [10.1093/jamiaopen/ooz046](https://doi.org/10.1093/jamiaopen/ooz046)] [Medline: [32734155](https://pubmed.ncbi.nlm.nih.gov/32734155/)]
11. Waring J, Lindvall C, Umeton R. Automated machine learning: Review of the state-of-the-art and opportunities for healthcare. *Artif Intell Med* 2020 Apr;104:101822 [FREE Full text] [doi: [10.1016/j.artmed.2020.101822](https://doi.org/10.1016/j.artmed.2020.101822)] [Medline: [32499001](https://pubmed.ncbi.nlm.nih.gov/32499001/)]
12. Korngiebel DM, Mooney SD. Considering the possibilities and pitfalls of Generative Pre-trained Transformer 3 (GPT-3) in healthcare delivery. *NPJ Digit Med* 2021 Jun 03;4(1):93 [FREE Full text] [doi: [10.1038/s41746-021-00464-x](https://doi.org/10.1038/s41746-021-00464-x)] [Medline: [34083689](https://pubmed.ncbi.nlm.nih.gov/34083689/)]
13. Miller K. When algorithmic fairness fixes fail: The case for keeping humans in the loop. Stanford University. 2020 Nov 02. URL: <https://hai.stanford.edu/news/when-algorithmic-fairness-fixes-fail-case-keeping-humans-loop> [accessed 2021-12-14]
14. Lucy L, Bamman D. Gender and representation bias in GPT-3 generated stories. In: *Proceedings of the Third Workshop on Narrative Understanding*. 2021 Jun Presented at: Third Workshop on Narrative Understanding; June 11, 2021; Virtual p. 48-55 URL: <https://aclanthology.org/2021.nuse-1.5.pdf> [doi: [10.18653/v1/2021.nuse-1.5](https://doi.org/10.18653/v1/2021.nuse-1.5)]
15. Bender EM, Gebru T, McMillan-Major A, Shmitchell S. On the dangers of stochastic parrots: Can language models be too big? 2021 Mar Presented at: FAccT '21: 2021 ACM Conference on Fairness, Accountability, and Transparency; March 3-10, 2021; Canada. [doi: [10.1145/3442188.3445922](https://doi.org/10.1145/3442188.3445922)]
16. McGuffie K, Newhouse A. The radicalization risks of GPT-3 and advanced neural language models. arXiv. Preprint posted online on September 15, 2020 [FREE Full text]
17. Daws R. Medical chatbot using OpenAI's GPT-3 told a fake patient to kill themselves. *AI News*. 2020 Oct 28. URL: <https://artificialintelligence-news.com/2020/10/28/medical-chatbot-openai-gpt3-patient-kill-themselves/> [accessed 2022-02-01]
18. FitzGerald C, Hurst S. Implicit bias in healthcare professionals: a systematic review. *BMC Med Ethics* 2017 Mar 01;18(1):19 [FREE Full text] [doi: [10.1186/s12910-017-0179-8](https://doi.org/10.1186/s12910-017-0179-8)] [Medline: [28249596](https://pubmed.ncbi.nlm.nih.gov/28249596/)]
19. Papineni K, Roukos S, Ward T, Zhu WJ. Bleu: a method for automatic evaluation of machine translation. 2002 Jul Presented at: 40th Annual Meeting of the Association for Computational Linguistics; July 7-12, 2002; Philadelphia, Pennsylvania, USA URL: <https://www.aclweb.org/anthology/P02-1040.pdf> [doi: [10.3115/1073083.1073135](https://doi.org/10.3115/1073083.1073135)]
20. Banerjee S, Lavie A. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. 2005 Jun Presented at: ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization; June 29, 2005; Ann Arbor, Michigan URL: <https://www.aclweb.org/anthology/W05-0909.pdf> [doi: [10.3115/1626355.1626389](https://doi.org/10.3115/1626355.1626389)]
21. Celikyilmaz A, Clark E, Gao J. Evaluation of text generation: A survey. arXiv. Preprint posted online on May 18, 2021 [FREE Full text]

22. Jiang R. Introducing new Alexa healthcare skills. Amazon Alexa. 2019 Apr 04. URL: <https://developer.amazon.com/blogs/alexa/post/ff33dbc7-6cf5-4db8-b203-99144a251a21/introducing-new-alexa-healthcare-skills> [accessed 2022-02-01]
23. McGraw D, Mandl KD. Privacy protections to encourage use of health-relevant digital data in a learning health system. NPJ Digit Med 2021 Jan 04;4(1):2 [FREE Full text] [doi: [10.1038/s41746-020-00362-8](https://doi.org/10.1038/s41746-020-00362-8)] [Medline: [33398052](https://pubmed.ncbi.nlm.nih.gov/33398052/)]
24. Artificial intelligence and machine learning in Software as a Medical Device. U.S. Food & Drug Administration. URL: <https://www.fda.gov/medical-devices/software-medical-device-samd/artificial-intelligence-and-machine-learning-software-medical-device> [accessed 2021-12-14]
25. Patient trust must come at the top of researchers' priority list. Nat Med 2020 Mar;26(3):301. [doi: [10.1038/s41591-020-0813-8](https://doi.org/10.1038/s41591-020-0813-8)] [Medline: [32161404](https://pubmed.ncbi.nlm.nih.gov/32161404/)]
26. Brockman G, Murati M, Welinder P, OpenAI. OpenAI API. OpenAI. 2020 Jun 11. URL: <https://openai.com/blog/openai-api/> [accessed 2021-12-14]
27. Romero A. Can't access GPT-3? Here's GPT-J — its open-source cousin. Towards Data Science. 2021 Jun 24. URL: <https://towardsdatascience.com/cant-access-gpt-3-here-s-gpt-j-its-open-source-cousin-8af86a638b11> [accessed 2021-12-14]
28. Sadilek A, Liu L, Nguyen D, Kamruzzaman M, Serghiou S, Rader B, et al. Privacy-first health research with federated learning. NPJ Digit Med 2021 Sep 07;4(1):132 [FREE Full text] [doi: [10.1038/s41746-021-00489-2](https://doi.org/10.1038/s41746-021-00489-2)] [Medline: [34493770](https://pubmed.ncbi.nlm.nih.gov/34493770/)]

## Abbreviations

**AI:** artificial intelligence

**BLEU:** Bilingual Evaluation Understudy

**EHR:** electronic health record

**GPT-3:** generative pretrained transformer 3

**HIPAA:** Health Insurance Portability and Accountability Act

**METEOR:** Metric for Evaluation of Translation With Explicit Ordering

**PCORI:** Patient-Centered Outcomes Research Institute

*Edited by C Lovis; submitted 12.08.21; peer-reviewed by A Trojan, T Caze, G Carot-Sans, J Ayre, W Zhang; comments to author 24.10.21; revised version received 14.12.21; accepted 09.01.22; published 10.02.22.*

*Please cite as:*

*Sezgin E, Sirrianni J, Linwood SL*

*Operationalizing and Implementing Pretrained, Large Artificial Intelligence Linguistic Models in the US Health Care System: Outlook of Generative Pretrained Transformer 3 (GPT-3) as a Service Model*

*JMIR Med Inform 2022;10(2):e32875*

*URL: <https://medinform.jmir.org/2022/2/e32875>*

*doi: [10.2196/32875](https://doi.org/10.2196/32875)*

*PMID: [35142635](https://pubmed.ncbi.nlm.nih.gov/35142635/)*

©Emre Sezgin, Joseph Sirrianni, Simon L Linwood. Originally published in JMIR Medical Informatics (<https://medinform.jmir.org>), 10.02.2022. This is an open-access article distributed under the terms of the Creative Commons Attribution License (<https://creativecommons.org/licenses/by/4.0/>), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Medical Informatics, is properly cited. The complete bibliographic information, a link to the original publication on <https://medinform.jmir.org/>, as well as this copyright and license information must be included.

Viewpoint

# A Free, Open-Source, Offline Digital Health System for Refugee Care

Henry Ashworth<sup>1,2</sup>, MPH; Senan Ebrahim<sup>2</sup>, MD, PhD; Hassaan Ebrahim<sup>2,3</sup>; Zahra Bhaiwala<sup>2</sup>; Michael Chilazi<sup>2,4</sup>, MD

<sup>1</sup>Department of Medicine, Harvard Medical School, Boston, MA, United States

<sup>2</sup>Hikma Health, San Jose, CA, United States

<sup>3</sup>Department of Public Policy, Harvard Kennedy School, Boston, MA, United States

<sup>4</sup>Johns Hopkins Hospital, Baltimore, MD, United States

**Corresponding Author:**

Henry Ashworth, MPH  
Department of Medicine  
Harvard Medical School  
25 Shattuck Street  
Boston, MA, 02115  
United States  
Phone: 1 8052152433  
Email: [hcashwor@gmail.com](mailto:hcashwor@gmail.com)

## Abstract

**Background:** Rise of conflict, extreme weather events, and pandemics have led to larger displaced populations worldwide. Displaced populations have unique acute and chronic health needs that must be met by low-resource health systems. Electronic health records (EHRs) have been shown to improve health outcomes in displaced populations, but need to be adapted to meet the constraints of these health systems.

**Objective:** The aim of this viewpoint is to describe the development and deployment of an EHR designed to care for displaced populations in low-resource settings.

**Methods:** Using a human-centered design approach, we conducted in-depth interviews and focus groups with patients, health care providers, and administrators in Lebanon and Jordan to identify the essential EHR features. These features, including modular workflows, multilingual interfaces, and offline-first capabilities, led to the development of the Hikma Health EHR, which has been deployed in Lebanon and Nicaragua.

**Results:** We report the successes and challenges from 12 months of Hikma Health EHR deployment in a mobile clinic providing care to Syrian refugees in Bekaa Valley, Lebanon. Successes include the EHR's ability to (1) increase clinical efficacy by providing detailed patient records, (2) be adaptable to the threats of COVID-19, and (3) improve organizational planning. Lessons learned include technical fixes to methods of identifying patients through name or their medical record ID.

**Conclusions:** As the number of displaced people continues to rise globally, it is imperative that solutions are created to help maximize the health care they receive. Free, open-sourced, and adaptable EHRs can enable organizations to better provide for displaced populations.

(*JMIR Med Inform* 2022;10(2):e33848) doi:[10.2196/33848](https://doi.org/10.2196/33848)

**KEYWORDS**

electronic health record; mHealth; refugee; displaced population; digital health; COVID-19; health care

## Introduction

There are over 80 million displaced people worldwide, and this number is projected to rise with increasing rates of natural disasters, conflict, and infectious disease outbreaks [1]. Over the last year, natural disasters and conflicts have been compounded by the impact of the COVID-19 pandemic.

Displaced populations have been significantly affected due to limited ability to implement social distancing measures and minimal access to vaccines [2]. At baseline, displaced populations face unique health threats, ranging from violence to food insecurity, infectious diseases, and exacerbation of underlying chronic conditions [3].

Limited resources and continual displacement create unique challenges for effective health care delivery to refugees, particularly when superimposed on the burdens of chronic diseases such as diabetes and hypertension [2]. When it comes to meeting these challenges, health care organizations caring for displaced populations often operate without formal systems for maintaining records of patient information. The lack of formal record-keeping can create even more challenges in establishing regular and consistent care [3]. This lack of consistency particularly impacts the continuity of care for displaced persons with chronic diseases. Overall, disjointed systems lead to an increased patient burden and perpetuate poor health outcomes [4].

A systematic review showed that general health records improved health outcomes in refugee populations [5]. Although electronic health record (EHR) systems have traditionally been built for high-resource settings optimized for billing, they have the unique ability of being adaptable to meet the needs of health care settings serving displaced populations. Previous research on implementing EHRs in displaced populations has shown that they lead to better patient outcomes through tracking of disease markers, increasing provider adherence to guidelines, and increasing patient adherence [4,6-10]. However, there are important barriers to effectively implementing EHRs in displaced population settings. In particular, EHRs need to be adaptable, and providers need to be engaged to drive overall uptake and success [6,7,11].

With this foundation in mind, we started a nonprofit organization, Hikma Health, to develop an EHR system to meet these needs and advance the care provided for displaced populations. Hikma Health was first started as an initiative from the Massachusetts Institute of Technology (MIT) Media Lab's 2017 Refugee Learning Accelerator. During a visit with refugee youth and their caregivers in Amman, Jordan, we realized the need for a free and open-source EHR system to provide continuity in their care. We incorporated as a 501(c)(3) nonprofit organization in California. Our initial seed funding came from the Harvard Business School New Venture Competition, and we have since grown with the support of private philanthropic foundations.

## Methods

### EHR Design Methodology

We refocused EHR development by employing a human-centered design approach that has been used successfully in global health technology, particularly for chronic disease management [12]. We started by conducting 30 patient and 12 provider interviews in Jordan, Lebanon, Turkey, and Greece. These interviews included a variety of displaced population

health care settings, including mobile clinics, multispecialty clinics, and hospitals, to identify current gaps in documentation and needs to improve care through desired workflows. Additionally, we collected documentation templates reflecting providers' then-current paper-based health record-keeping in Lebanon. We observed the clinician practice at five clinical practices in both rural and urban settings in Lebanon and Jordan. We also conducted three focus groups on the preliminary design concept with providers of different specialties, including internal medicine, nursing, neurology, cardiology, pediatrics, dentistry, mental health, and social work. The results from the interviews and focus groups were analyzed using a framework analysis. We designed the architecture of the Hikma Health system and prototyped it with the early support of these providers and patients.

### Hikma Health Design Features

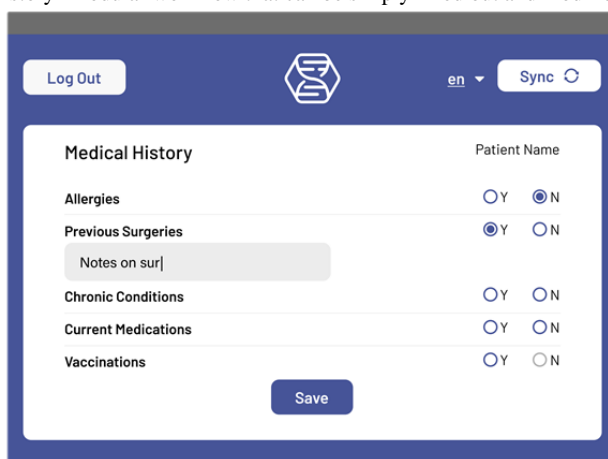
Through our human-centered design process, we identified three essential features for an effective digital health system in low-resource settings: modular workflows, multilingual interfaces, and offline-first capabilities. We then designed the Hikma Health app as an offline-first multilingual mobile EHR system with 32 modular workflows and sensible defaults for the global care of mobile populations, including a clinically validated COVID-19 screening tool. Although first-generation EHRs were limited by technical constraints [13], modern EHRs such as the Hikma Health system incorporate modular workflows. This modularity is essential to appropriately provisioning the app based on local clinical practice, such as consulting specialty care services or tracking prescriptions over time [14]. Modular workflows refer to distinct premade documentation forms that can be included in a particular EHR deployment based on services available at that clinic. Working within this modular framework enables clinics to deploy a functional system rapidly by provisioning the relevant workflows as plug-and-play modules (Figure 1). Each module is further modifiable, offering full customizability as needed to optimize care (Figure 2).

In light of the well-documented challenges for multilingual care practices in refugee care settings, we built the Hikma Health app to fully support English, Arabic, and Spanish with support for instantaneous translation for multilingual provider teams. In addition, recent advances in database technology and smartphone hardware enabled us to provision the Hikma Health system as a functional offline-first app with automatic data synchronization to a central server whenever a mobile device is connected. An offline-first capability continues to be essential in many remote areas where displaced populations receive care; however, most EHR systems designed for the developed world do not have this capability.

**Figure 1.** New Visit dashboard showing different modular workflows that can be modified for a clinic’s needs.



**Figure 2.** An example of the “medical history” modular workflow that can be simply filled out and modified.



**Sustainability**

Hikma Health initially supported the deployment of the Hikma Health system directly by providing cloud storage and mobile devices, employing US-based technical staff to customize the deployment, and supporting local engineers in-country to take over the deployment. This model was optimal for the early development and optimization of the software system in 2018-2020. In 2021, we launched the Year of Migrant Health

program to support independent deployment of the stable release of the Hikma Health system at a wider, global scale. In this grant program, we are granting up to US \$25,000 per clinical organization to offset costs of independent deployment, including cloud storage, mobile devices, and engineering staff costs. We believe the Year of Migrant Health deployment model will expand the use of the Hikma Health system, strengthen its open-source modular workflow library as developers contribute

globally, and demonstrate a sustainable means for clinics to autonomously deploy their EHRs.

## Deployment

The open-source Hikma Health system is freely available to organizations with the technical capacity to self-deploy the platform on a cloud server. However, many clinical organizations working in low-resource settings lack the technical expertise and resources to self-deploy. Therefore, our technical team works directly with clinical partners to both manage deployment and customize our generic EHR system to fit their specific medical workflows.

The Hikma Health system is currently in use by two clinical partners, Endless Medical Advantage (EMA) and Nueva Vida Clinic. EMA is a mobile health clinic based in Bekaa Valley in Lebanon that travels throughout the region serving hundreds of informal and dispersed Syrian refugee camps, as well as vulnerable local communities. EMA has been actively using the Hikma Health system since September 2020. Nueva Vida Clinic in Ciudad Sandino, Nicaragua, is a primary health care facility that provides free and subsidized care to rural and low-income communities, including migratory populations. Nueva Vida Clinic just started actively using the Hikma Health system since September 2021. We have been assessing the strengths and limitations of the Hikma Health system through monthly quality assurance and quality improvement meetings with our partners.

## Ethical Considerations

A formal ethical approval for this study was not sought as it was a general inquiry concerning EHR perspectives and by the Harvard Longwood Campus Institutional Review Board it did not meet the US federal definition for research.

## Results

### Reported Outcomes

In the 12 months that EMA has been using the Hikma Health system, they have been able to transition from a paper record-based system of patient health data collection to a digital system. EMA health care workers have reported that using the Hikma Health system has increased the efficiency with which they are able to collect and access patient health information, particularly in the field while seeing patients. Previously, EMA reported that they had been unable to see past the visit history of a particular patient, but with the Hikma Health system were able to view all previously collected health information, including past diagnoses and prescriptions that were of particular importance. Additionally, EMA physicians and the administration have specified that the offline functionality of the system was essential for their operations, given the limited network connectivity throughout Bekaa Valley.

The customizability of the Hikma Health system also enabled continuous improvement to the medical forms and workflows throughout the past year, enabling the Hikma Health team of developers to easily adapt the system to fit EMA's dynamic needs. For example, EMA was able to deploy a screening

module for COVID-19 to easily identify high-risk patients for isolation and testing.

Prior to using the Hikma Health system, EMA administrative staff would have to manually process paper records to compile key organizational metrics of the organization, such as total patient visits, prevalence of particular diagnoses, and number of medications prescribed. These organizational metrics are essential to the clinic's programmatic efficiency and operations. With the Hikma Health system's administrative dashboard, which allows administrative staff to easily export all of the clinic's patient data, administrative staff reported that they were able to track outcomes of the clinic with greater ease compared to the paper-based system.

As Nueva Vida Clinic has just started to use the Hikma Health system, we do not have any reported results from the field. However, important lessons were learned from the customization of the Hikma Health system to the needs of Nueva Vida Clinic.

## Technical Issues and Lessons Learned

One of the challenges that arose during the deployment with EMA surrounded the ability to easily search and filter for individual patients. Because many Arabic names are transliterated in multiple ways with different spellings (ie, Muhammad or Muhamad), clinician users were having a hard time finding the correct identity in an efficient manner. Although they were able to find patients using additional search parameters such as date of birth, phone number, or hometown, clinician users expressed frustration at the challenges with using a first-name search for many patients. In response, we implemented a new feature within the Hikma Health system that allowed for "fuzzy searches," which would include all variations of the spelling of the most commonly occurring names within the population. Although this feature required manually compiling this list of names and possible spellings, upon implementation, it dramatically improved the efficiency of the search feature within the field.

Another technical challenge we encountered during the deployment with EMA was an issue with the syncing feature of the system when certain fields within the patient registration were missing. The backend database of the system generated a unique user identification number as a combination of digits including a patient's date of birth. If during patient registration, the date of birth was not collected by the clinician user, the unique user identification number would be unable to be generated, creating an error when trying to sync the data. This issue prevented successful synchronization for multiple days before being noticed by the clinical team. In response, we pushed out a software fix that resolved the issue of the unique user identification number not being created for patients with a missing date of birth. Furthermore, we created an alert with the mobile app that notified clinician users of the status of the sync, either confirming its success or flagging its failure.

## Discussion

### Principal Findings

As natural disasters, conflicts, and infectious disease outbreaks continue to increase the number of people displaced worldwide,

innovations are needed to meet their needs and the limitations of settings in which health care is provided. EHRs are one possible solution, yet they must be adaptable and implemented with strong provider uptake and buy-in [6,7,11]. By taking a user-centered design approach, we have created an EHR that meets many of these needs, including offline capabilities for areas without internet connection, modular workflows to simplify the user interface, and multilingual capability to increase accessibility. Compared to other EHRs, our system is fully open-sourced [7,10,15]. This allows any use with the support of a software engineer to adapt the Hikma Health EHR platform to meet a clinic's needs. Our EHR also has a few limitations, including no direct way for patients to carry a copy of their own medical record, an innovative feature introduced by the Sijilli EHR created by Epic [15]. Although the Hikma Health EHR platform does not contain every element to meet a health system's needs out of the box, this adaptability is a central feature allowing for customization.

From our informal results to date, we have seen the Hikma Health system meet the needs of two different health care organizations. In the remote areas of Bekaa Valley to rural Nicaragua, our partners have been able to adapt the Hikma Health system to meet the constraints of their environment while still providing for their patients. To continue to ensure sustainability of the Hikma Health system, we support the hiring and training of in-country software engineers to address bugs or updates our partners require.

The biggest limitation of this system is that it may not be perfectly suited for every clinic once downloaded. As each

health care setting caring for displaced populations is different, we cannot create an EHR to meet every need. Additionally, it does take time and resources for a health care setting to actually set up an EHR. These resources include hardware and technical personnel. As previously mentioned, an EHR is not a solution within itself. Finally, to be successful, any EHR requires engagement of all its users from health care providers to administrators. Therefore, each EHR, including our own, requires early and continual engagement to ensure long-term success.

Looking to the future, it is clear from research to date that EHR innovations for displaced populations offer a means to improve care [4-10]. Future innovations should continue to take user-centered design approaches and consider such factors as provider-tailored modular workflows and offline capabilities. Additionally, future innovations should consider ways to expand patient access to their medical records, ease data transfer across systems or organizations, integrate diagnostic technologies, and decrease technical barriers to implementation.

## Conclusions

As displaced people around the world continue to face the COVID-19 pandemic and other health care challenges, it is imperative that adaptable EHR solutions are developed to meet their specific health needs. The Hikma Health system is a free EHR optimized for mobile, offline use that has been designed with the user in mind to meet the needs of displaced patients. By making the system free and open source, we aim to enhance every organization's capacity to provide better care for displaced populations worldwide.

---

## Acknowledgments

The Hikma Health EHR was developed with financial support from Fast Forward, MassChallenge, the Gerson Lehrman Group, HP Enterprise, Google.org, BlackRock, the Harvard Business School New Ventures Competition, the MIT \$100K Competition, the MIT Media Lab, the MIT Sandbox Innovation Fund Program, and the Robert Wood Johnson Foundation.

---

## Conflicts of Interest

All authors are volunteers for Hikma Health, a 501(c)(3) nonprofit organization. None of the authors have ever received any financial compensation for their work.

---

## References

1. Global trends: forced displacement in 2019. United Nations High Commissioner for Refugees. 2019. URL: <https://www.unhcr.org/5ee200e37.pdf> [accessed 2021-06-10]
2. Alawa J, Zarei P, Khoshnood K. Evaluating the provision of health services and barriers to treatment for chronic diseases among Syrian refugees in Turkey: a review of literature and stakeholder interviews. *Int J Environ Res Public Health* 2019 Jul 25;16(15):2660 [FREE Full text] [doi: [10.3390/ijerph16152660](https://doi.org/10.3390/ijerph16152660)] [Medline: [31349639](https://pubmed.ncbi.nlm.nih.gov/31349639/)]
3. Doocy S, Lyles E, Robertson T, Akhu-Zaheya L, Oweis A, Burnham G. Prevalence and care-seeking for chronic diseases among Syrian refugees in Jordan. *BMC Public Health* 2015 Oct 31;15:1097 [FREE Full text] [doi: [10.1186/s12889-015-2429-3](https://doi.org/10.1186/s12889-015-2429-3)] [Medline: [26521231](https://pubmed.ncbi.nlm.nih.gov/26521231/)]
4. Doocy S, Paik KE, Lyles E, Hei Tam H, Fahed Z, Winkler E, et al. Guidelines and mHealth to improve quality of hypertension and type 2 diabetes care for vulnerable populations in Lebanon: longitudinal cohort study. *JMIR Mhealth Uhealth* 2017 Oct 18;5(10):e158 [FREE Full text] [doi: [10.2196/mhealth.7745](https://doi.org/10.2196/mhealth.7745)] [Medline: [29046266](https://pubmed.ncbi.nlm.nih.gov/29046266/)]
5. Chiesa V, Chiarenza A, Mosca D, Rechel B. Health records for migrants and refugees: a systematic review. *Health Policy* 2019 Sep;123(9):888-900 [FREE Full text] [doi: [10.1016/j.healthpol.2019.07.018](https://doi.org/10.1016/j.healthpol.2019.07.018)] [Medline: [31439455](https://pubmed.ncbi.nlm.nih.gov/31439455/)]

6. Doocy S, Lyles E, Akhu-Zaheya L, Oweis A, Al Ward N, Burton A. Health service utilization among Syrian refugees with chronic health conditions in Jordan. *PLoS One* 2016;11(4):e0150088 [FREE Full text] [doi: [10.1371/journal.pone.0150088](https://doi.org/10.1371/journal.pone.0150088)] [Medline: [27073930](https://pubmed.ncbi.nlm.nih.gov/27073930/)]
7. Doocy S, Paik K, Lyles E, Tam HH, Fahed Z, Winkler E, et al. Pilot testing and implementation of a mHealth tool for non-communicable diseases in a humanitarian setting. *PLoS Curr* 2017 Jun 05;9:ecurrents.dis.e98c648aac93797b1996a37de099be74. [doi: [10.1371/currents.dis.e98c648aac93797b1996a37de099be74](https://doi.org/10.1371/currents.dis.e98c648aac93797b1996a37de099be74)] [Medline: [28744410](https://pubmed.ncbi.nlm.nih.gov/28744410/)]
8. Khader A, Farajallah L, Shahin Y, Hababeh M, Abu-Zayed I, Kochi A, et al. Cohort monitoring of persons with hypertension: an illustrated example from a primary healthcare clinic for Palestine refugees in Jordan. *Trop Med Int Health* 2012 Sep;17(9):1163-1170. [doi: [10.1111/j.1365-3156.2012.03048.x](https://doi.org/10.1111/j.1365-3156.2012.03048.x)] [Medline: [22845700](https://pubmed.ncbi.nlm.nih.gov/22845700/)]
9. Khader A, Ballout G, Shahin Y, Hababeh M, Farajallah L, Zeidan W, et al. Diabetes mellitus and treatment outcomes in Palestine refugees in UNRWA primary health care clinics in Jordan. *Public Health Action* 2013 Dec 21;3(4):259-264 [FREE Full text] [doi: [10.5588/pha.13.0083](https://doi.org/10.5588/pha.13.0083)] [Medline: [26393043](https://pubmed.ncbi.nlm.nih.gov/26393043/)]
10. Khader A, Ballout G, Shahin Y, Hababeh M, Farajallah L, Zeidan W, et al. Treatment outcomes in a cohort of Palestine refugees with diabetes mellitus followed through use of E-Health over 3 years in Jordan. *Trop Med Int Health* 2014 Feb;19(2):219-223. [doi: [10.1111/tmi.12241](https://doi.org/10.1111/tmi.12241)] [Medline: [24341942](https://pubmed.ncbi.nlm.nih.gov/24341942/)]
11. Saleh S, Farah A, Dimassi H, El Arnaout N, Constantin J, Osman M, et al. Using mobile health to enhance outcomes of noncommunicable diseases care in rural settings and refugee camps: randomized controlled trial. *JMIR Mhealth Uhealth* 2018 Jul 13;6(7):e137 [FREE Full text] [doi: [10.2196/mhealth.8146](https://doi.org/10.2196/mhealth.8146)] [Medline: [30006326](https://pubmed.ncbi.nlm.nih.gov/30006326/)]
12. Holeman I, Kane D. Human-centered design for global health equity. *Inf Technol Dev* 2019 Sep 29;26(3):477-505 [FREE Full text] [doi: [10.1080/02681102.2019.1667289](https://doi.org/10.1080/02681102.2019.1667289)] [Medline: [32982007](https://pubmed.ncbi.nlm.nih.gov/32982007/)]
13. Roland L, Sanner T, Aanestad M. Flexibility in EHR ecosystems: five integration strategies and their trade-offs. 2017 Presented at: Norsk konferanse for organisasjoners bruk av IT; November 27-29; Oslo.
14. Collins DR, Jobanputra K, Frost T, Muhammed S, Ward A, Shafei AA, et al. Cardiovascular disease risk and prevention amongst Syrian refugees: mixed methods study of Médecins Sans Frontières programme in Jordan. *Confl Health* 2017;11:14. [doi: [10.1186/s13031-017-0115-z](https://doi.org/10.1186/s13031-017-0115-z)] [Medline: [28725259](https://pubmed.ncbi.nlm.nih.gov/28725259/)]
15. Saleh S, El Arnaout N, Abdouni L, Jammoul Z, Hachach N, Dasgupta A. Sijilli: a scalable model of cloud-based electronic health records for migrating populations in low-resource settings. *J Med Internet Res* 2020 Aug 13;22(8):e18183 [FREE Full text] [doi: [10.2196/18183](https://doi.org/10.2196/18183)] [Medline: [32788145](https://pubmed.ncbi.nlm.nih.gov/32788145/)]

## Abbreviations

**EHR:** electronic health record

**EMA:** Endless Medical Advantage

**MIT:** Massachusetts Institute of Technology

*Edited by C Lovis; submitted 26.09.21; peer-reviewed by L de Witte, N Slabyj; comments to author 20.12.21; revised version received 27.12.21; accepted 02.01.22; published 11.02.22.*

*Please cite as:*

Ashworth H, Ebrahim S, Ebrahim H, Bhaiwala Z, Chilazi M

A Free, Open-Source, Offline Digital Health System for Refugee Care

*JMIR Med Inform* 2022;10(2):e33848

URL: <https://medinform.jmir.org/2022/2/e33848>

doi: [10.2196/33848](https://doi.org/10.2196/33848)

PMID: [35147509](https://pubmed.ncbi.nlm.nih.gov/35147509/)

©Henry Ashworth, Senan Ebrahim, Hassaan Ebrahim, Zahra Bhaiwala, Michael Chilazi. Originally published in *JMIR Medical Informatics* (<https://medinform.jmir.org/>), 11.02.2022. This is an open-access article distributed under the terms of the Creative Commons Attribution License (<https://creativecommons.org/licenses/by/4.0/>), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in *JMIR Medical Informatics*, is properly cited. The complete bibliographic information, a link to the original publication on <https://medinform.jmir.org/>, as well as this copyright and license information must be included.



Original Paper

# Evaluation of Natural Language Processing for the Identification of Crohn Disease–Related Variables in Spanish Electronic Health Records: A Validation Study for the PREMONITION-CD Project

Carmen Montoto<sup>1</sup>, MD, PhD; Javier P Gisbert<sup>2,3,4,5</sup>, MD, PhD; Iván Guerra<sup>6</sup>, MD, PhD; Rocío Plaza<sup>7</sup>, MD; Ramón Pajares Villarroya<sup>8</sup>, MD; Luis Moreno Almazán<sup>9</sup>, MD; María Del Carmen López Martín<sup>10</sup>, MD; Mercedes Domínguez Antonaya<sup>11</sup>, MD; Isabel Vera Mendoza<sup>12</sup>, MD, PhD; Jesús Aparicio<sup>1</sup>, PhD; Vicente Martínez<sup>1</sup>, MD, PhD; Ignacio Tagarro<sup>1</sup>, PhD; Alonso Fernandez-Nistal<sup>1</sup>, PhD; Lea Canales<sup>13</sup>, PhD; Sebastian Menke<sup>14</sup>, PhD; Fernando Gomollón<sup>15,16,17,18</sup>, MD, PhD; PREMONITION-CD Study Group<sup>19</sup>

<sup>1</sup>Takeda Farmacéutica España S.A., Madrid, Spain

<sup>2</sup>Hospital Universitario de La Princesa, Madrid, Spain

<sup>3</sup>Instituto de Investigación Sanitaria Princesa (IIS-IP), Madrid, Spain

<sup>4</sup>Universidad Autónoma de Madrid, Madrid, Spain

<sup>5</sup>Centro de Investigación Biomédica en Red de Enfermedades Hepáticas y Digestivas (CIBEREHD), Madrid, Spain

<sup>6</sup>Hospital Universitario de Fuenlabrada, Madrid, Spain

<sup>7</sup>Hospital Universitario Infanta Leonor, Madrid, Spain

<sup>8</sup>Hospital Universitario Infanta Sofía, Madrid, Spain

<sup>9</sup>Hospital Universitario HM Montepríncipe, Madrid, Spain

<sup>10</sup>Hospital Universitario Infanta Elena, Madrid, Spain

<sup>11</sup>Hospital Universitario Rey Juan Carlos, Madrid, Spain

<sup>12</sup>Hospital Universitario Puerta de Hierro Majadahonda, Madrid, Spain

<sup>13</sup>Department of Software and Computing System, University of Alicante, Alicante, Spain

<sup>14</sup>MedSavana SL, Madrid, Spain

<sup>15</sup>Hospital Clínico Universitario Lozano Blesa, Zaragoza, Spain

<sup>16</sup>Instituto de Investigación Sanitaria Aragón (IISA), Zaragoza, Spain

<sup>17</sup>Universidad de Zaragoza, Zaragoza, Spain

<sup>18</sup>Centro de Investigación Biomédica en Red de Enfermedades Hepáticas y Digestivas (CIBEREHD), Zaragoza, Spain

<sup>19</sup>See Acknowledgements

**Corresponding Author:**

Carmen Montoto, MD, PhD

Takeda Farmacéutica España S.A.

Edificio Torre Europa

Paseo de la Castellana, 95

Madrid, 28046

Spain

Phone: 34 917904222

Email: [Carmen.montoto@takeda.com](mailto:Carmen.montoto@takeda.com)

## Abstract

**Background:** The exploration of clinically relevant information in the free text of electronic health records (EHRs) holds the potential to positively impact clinical practice as well as knowledge regarding Crohn disease (CD), an inflammatory bowel disease that may affect any segment of the gastrointestinal tract. The EHRead technology, a clinical natural language processing (cNLP) system, was designed to detect and extract clinical information from narratives in the clinical notes contained in EHRs.

**Objective:** The aim of this study is to validate the performance of the EHRead technology in identifying information of patients with CD.

**Methods:** We used the EHRead technology to explore and extract CD-related clinical information from EHRs. To validate this tool, we compared the output of the EHRead technology with a manually curated gold standard to assess the quality of our cNLP system in detecting records containing any reference to CD and its related variables.

**Results:** The validation metrics for the main variable (CD) were a precision of 0.88, a recall of 0.98, and an F1 score of 0.93. Regarding the secondary variables, we obtained a precision of 0.91, a recall of 0.71, and an F1 score of 0.80 for CD flare, while for the variable vedolizumab (treatment), a precision, recall, and F1 score of 0.86, 0.94, and 0.90 were obtained, respectively.

**Conclusions:** This evaluation demonstrates the ability of the EHRead technology to identify patients with CD and their related variables from the free text of EHRs. To the best of our knowledge, this study is the first to use a cNLP system for the identification of CD in EHRs written in Spanish.

(*JMIR Med Inform* 2022;10(2):e30345) doi:[10.2196/30345](https://doi.org/10.2196/30345)

## KEYWORDS

natural language processing; linguistic validation; artificial intelligence; electronic health records; Crohn disease; inflammatory bowel disease

## Introduction

Crohn disease (CD) is a chronic inflammatory bowel disease (IBD) that leads to lesions in different sites along the length of the gastrointestinal tract and, occasionally, in other extraintestinal locations such as skin, eyes, joints, mouth, and the hepatobiliary system [1]. Symptoms (including abdominal pain, diarrhea, fever, and weight loss) evolve in a relapsing and remitting manner, leading to bowel damage and disability. CD is considered to be a heterogeneous disorder with a multifactorial etiology, in which genetics and environmental factors interact to manifest the disease [2]. Although most patients with CD are diagnosed with an inflammatory phenotype, about half of them do require surgeries derived from complications such as strictures, fistulas, or abscesses [3].

Over the last years, most health care institutions have moved away from paper clinical records toward electronic health records (EHRs) in which patients' longitudinal medical information is stored [4]. Since then, large volumes of digitalized real-world clinical data have been generated at exponential rates. Although some clinical data contained in the EHRs are stored in structured fields, the majority of the relevant clinical information appears embedded in the free-text narratives written down by health professionals [5].

The area of computer science dedicated to the analysis and representation of naturally occurring texts (written or oral) [6] is called natural language processing (NLP). One of the applications of NLP focuses on the extraction of information from free text captured in EHRs and is therefore referred to as clinical NLP (cNLP). So far, cNLP systems have been successfully applied for the extraction of relevant clinical information using approaches such as regular expressions or machine learning. As a result, the quantity and quality of data captured from the EHRs have substantially increased over recent years [7]. Although incorporating information from free text into case detection through NLP techniques improves research quality [8-10], one key challenge in this process is to ensure the validity of the results by assessing the detection performance.

In this context, as part of the PREMONITION-CD observational study, we aimed to assess the performance of the cNLP system *EHRead* technology [11-15] in identifying medical records that

contain mentions of CD and its related variables when compared to the detection performed by expert medical doctors. Because the manual review of free-text narratives is extremely time-consuming, valuable information routinely collected in clinical practice has largely remained unused for research purposes. Therefore, the validated automatic extraction of this information holds potential to advance our knowledge about CD and could have a positive impact in the management of these patients [16,17].

## Methods

### Ethics Approval and Consent to Participate

This study was conducted within the scope of the PREMONITION-CD project, a multicenter, retrospective study aimed at using NLP to detect free-text information in CD patients' EHRs. Before the start of data collection, the study was approved by the Spanish Ethics Committee, Agencia Española de Medicamentos y Productos Sanitarios, and the Madrid region Ethics Committee, Comité Ético de Investigación con Medicamentos Regional de la Comunidad de Madrid, with reference number IBD-5002 (May 2018). Approval from each of the hospitals participating in the study was also obtained. It was registered in *ClinicalTrials.gov* with the identifier number NCT03668249.

The study was conducted in compliance with legal and regulatory requirements and followed generally accepted research practices described in the ICH Guideline for Good Clinical Practice, the Declaration of Helsinki in its latest edition, Good Pharmacoepidemiology Practices, and applicable local regulations.

### Consent for Publication

In accordance with article 14.5 of the General Data Protection Regulation (GDPR), if obtaining consent is impossible or would involve a disproportionate effort, in particular for processing for archiving purposes in the public interest, scientific or historical research purposes, or statistical purposes, the study is subject to the conditions and safeguards referred to in Article 89.

Regarding Article 89 of the GDPR, processing in the public interest or scientific research purposes shall be subject to

appropriate safeguards and will not require consent from each of the data subjects, in accordance with the GDPR, for the rights and freedoms of the data subject.

**Availability of Data and Materials**

Due to the retrospective nature of the research, data analysis did not require consent from the data subjects. Therefore, supporting data is subject to strict confidentiality agreements with each participating hospital and cannot be made openly available.

**Data Source**

Data were collected from 8 hospitals of the Spanish National Healthcare Network from January 1, 2014, to December 31, 2018 (except for one participating site with electronic data available from 2013 to 2017).

**Study Design**

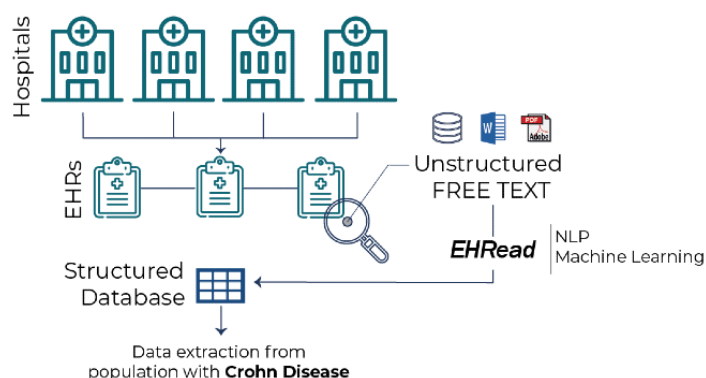
For this study, the assessed variables were CD, CD flare (a crucial variable for the characterization of the evolution of the

disease), and vedolizumab (a biologic drug indicated exclusively for the treatment of IBD). The variables included in this study were selected by the senior study committee based on the PREMONITION-CD overall study objectives. The variables were detected when written directly in the EHRs, without inferences or prior outcome definitions. The human annotations served the purpose of the creation of a gold standard to which the EHRead technology was compared.

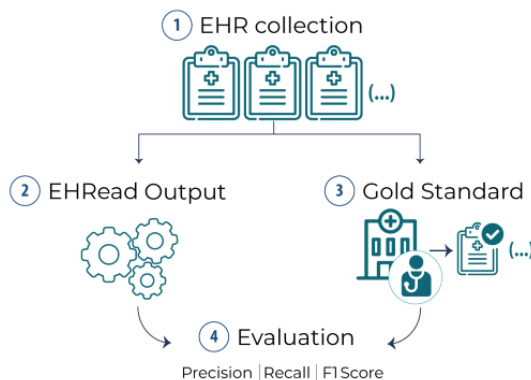
The EHRead technology is an NLP system designed to retrieve large amounts of biomedical information contained in EHRs [11-15] and convert the information into a structured representation (Figure 1).

To perform this study, we completed the following steps: EHR collection, processing using EHRead technology, creation of the gold standard data set, and comparison of both outputs using standard measures of performance (Figure 2).

**Figure 1.** Extracting and organizing unstructured clinical data into a structured database. The EHRead technology is a clinical NLP system that detects and extracts clinically relevant information contained in deidentified EHRs. The extracted information from participating sites is organized in a structured study database. From this database, patients that fulfill the study criteria based on the study inclusion and exclusion criteria make up the target population. In this case, clinical data from the population with a diagnosis of Crohn disease were used. EHR: electronic health record; NLP: natural language processing.



**Figure 2.** Linguistic evaluation process. To validate the output of the EHRead technology, a statistical comparison was performed between its output and a gold standard consisting of a subset of EHRs annotated by expert physicians. The validation metrics calculated are expressed in terms of precision, recall, and F1 score. See text for further details. EHR: electronic health record.



In the *EHR collection step*, a data set was selected that consisted of a sample collection of EHRs obtained primarily from the gastroenterology service (including consultation, hospitalization, and emergency reports), representing more than 3,900,000 patients. To obtain a representative data set, 100 records were

randomly selected from each of the 8 sites containing EHRs with and without CD-related information, amounting to a total of 800 clinical documents from 800 patients. Subsequently, all records were fully anonymized to meet legal and ethical requirements before they were annotated by physicians

(annotators) to generate a gold standard for each participating site (see sections about annotation process and gold standard).

In parallel to the annotation task carried out by physicians, the EHRead technology was applied on the free text of the same EHRs used to generate the gold standard (for more details see NLP System). By doing so, the performance of the EHRead technology could directly be compared to human performance in detection of CD and secondary variables.

In the final step of the evaluation, the performance of the EHRead technology was compared against the gold standard to validate the capacity of the technology in identifying records containing mentions of CD and its related variables. Therefore, both the detections of physicians and the EHRead technology were transformed into binaries (0 no detection, 1 detection) for each variable to calculate the performance metrics precision, recall, and F1 score using the library scikit-learn [18].

### NLP System

The main phases of the NLP system were the following:

- The section identification phase aims to detect the different parts of a clinical document, such as family medical history, physical exam, and treatment.
- The concept identification phase is when the system detects a medical concept. Specifically, the terminology considered by the EHRead technology is built upon SNOMED-CT (Systemized Nomenclature of Medicine–Clinical Terms), a leading platform of systematically organized and computer-readable collections of medical concepts. SNOMED-CT includes codes, concepts, synonyms, and definitions used in clinical documentation and is considered the most comprehensive terminology worldwide.
- The contextual information phase focuses on detecting the attributes of the already identified clinical terms within their textual context, both from an intention perspective (the term is either stated in an affirmative way or negated, or is part of a conjecture or opinion) and from a temporal perspective (current or historical).

### Annotation Process and Gold Standard

The manual revision of clinical texts was carried out by annotators specialized in gastroenterology. For the annotation task, guidelines were jointly created by internal NLP experts and clinical experts. They included the variables to be annotated in the free text, along with recommendations on how to solve uncertainties. Following these guidelines, specialists reviewed the free text of selected EHRs for the occurrence of the study variables to answer a set of yes/no questions: Does/did the patient have CD? Does the report state that the patient has had a flare? and Does the record state that the patient was treated with Vedolizumab? The second and third questions were only asked if the first one was affirmative, meaning that the patient did have CD before or at the time point of the hospital visit. The annotators were not allowed to respond with *yes* to any of the questions based on inference.

Of the 100 records selected per site, 15 were reviewed by two independent annotators to assess the interannotator agreement [19,20]. A low agreement indicates that the annotators had difficulties in linguistically identifying the relevant variables in the EHRs or that the guidelines are still inadequate in properly describing the annotation task [21]. Thus, the interannotator agreement serves as a control mechanism to check the reliability of the annotation and further to establish a target of performance for the NLP system. For this task, the annotators were not allowed to communicate with each other or share information regarding the annotation process to avoid bias. Once the annotations were finished, the interannotator agreement was calculated in terms of F1 score. Once the quality of annotations had been verified through the interannotator agreement and the disagreements had been resolved to build the final gold standard, one of the two physicians annotated the remaining 85% of clinical records to complete the gold standard.

### Statistical Analysis

The performance of the EHRead technology in identifying CD and its related variables was compared with the gold standard. The agreement between them was calculated using three metrics: *precision* (ie, positive predictive value), *recall* (ie, sensitivity), and their harmonic mean *F1 score* [21]. *Precision* is the indicator of the accuracy of information retrieved by the system, *recall* is the indicator of the amount of information the system retrieves, and *F1 score* conveys the balance between precision and recall. In addition to those metrics, we calculated the 95% CI for each aforementioned measure, since this provides information about the range in which the true value lies and thus how robust the metric is. The method used to calculate the 95% CIs is the Clopper-Pearson approach, one of the most common methods for calculating binomial 95% CIs.

### Results

The gold standard data set (N=800) consisted of 41.4% (n=331) medical records with CD, 21.3% (n=170) with CD flare, and 10% (n=83) with vedolizumab treatment. Table 1 shows the interannotator agreement F1 scores of the gold standard for each investigated variable per site.

The interannotator agreement values were higher than 0.8 for all comparisons, indicating an *almost perfect* agreement according to the Landis and Koch scale [19]. In addition, the overall agreement between all sites was *almost perfect* [22] for the three studied variables. The EHRead technology results in terms of *precision*, *recall*, and *F1 score* are shown in Table 2.

The detection of the main variable (ie, CD) achieved a *precision* of 0.88, a *recall* of 0.98, and an *F1 score* of 0.93. Regarding the secondary variables, CD flare obtained a *precision* of 0.91, a *recall* of 0.71, and an *F1 score* of 0.80, while the variable vedolizumab was detected at a *precision* of 0.86, a *recall* of 0.94, and an *F1 score* of 0.90.

**Table 1.** Interannotator agreement (F1 score) per participating site.

	F1 score		
	Crohn disease	Crohn disease flare	Vedolizumab
Site 1	0.93	0.86	1.00
Site 2	1.00	0.87	1.00
Site 3	1.00	1.00	1.00
Site 4	0.93	1.00	1.00
Site 5	0.93	0.83	1.00
Site 6	0.93	1.00	1.00
Site 7	1.00	1.00	1.00
Site 8	1.00	0.85	1.00
Average	0.97	0.93	1.00

**Table 2.** Performance of the EHRead technology.

Variable	Precision (95% CI)	Recall (95% CI)	F1 score (95% CI)
Crohn disease	0.88 (0.85-0.91)	0.98 (0.95-0.99)	0.93 (0.90-0.95)
Crohn disease flare	0.91 (0.85-0.95)	0.71 (0.63-0.77)	0.80 (0.72-0.85)
Vedolizumab	0.86 (0.76-0.93)	0.94 (0.86-0.98)	0.90 (0.81-0.96)

## Discussion

The evaluation presented here is part of the observational, retrospective PREMONITION-CD study, designed to characterize clinical and nonclinical variables of patients with CD. To the best of our knowledge, this is the first multicentric study using a cNLP system for the identification of prespecified CD-related variables from reports written in Spanish. The intrinsic characteristics of IBD and the current dilemmas associated with the medical management of affected patients present an opportunity for the implementation of big data research strategies. Artificial intelligence techniques complement current research efforts and might be key in disentangling the complexity of the disease [23] by allowing key patient-centered information to be retrieved and analyzed at a larger population scale. In turn, large CD/IBD data sets will enable the identification of clinical patterns, patient management, and predictors of disease that will ultimately improve patient care.

Although some clinical data is stored in structured fields of EHRs, the majority is contained in the narrative free text [4]. The automated extraction of these data using modern NLP techniques has a strikingly positive impact on clinical practice, since it enables the exploration of this valuable patient information at a scale that was not possible before. Here, we evaluated Savana's EHRead technology, a cNLP system designed to detect and extract clinically relevant information from the free text of EHRs [11-15], to identify CD reports from narrative clinical data.

In contrast to other research studies that applied NLP techniques on Spanish EHRs obtained from a single medical center [24,25], this study combined data from eight large hospitals, thereby providing robustness and enabling generalizability. The

capabilities of the EHRead technology allowed us to process a wide range of document types and to handle the different internal structures of clinical reports from the different participating sites. In addition, the inclusion of different sites enhanced the variability and richness of the language regarding the evaluated variables. Indeed, the variables evaluated were expressed in different ways across sites, including discrepancies in abbreviations or acronyms.

*F1 scores* higher than 0.80 for all interannotator agreements ensure that the gold standard met the criteria to serve as reference. In addition, our study demonstrates a good performance of the EHRead technology in identifying reports that contain mentions of CD and CD-related variables. We obtained *F1 scores* higher than 90% for the main variable and close to 80% for the remaining variables (Table 2). Despite the intrinsic heterogeneity of EHRs resulting from a variability in physicians, data collection sites, and record completeness, EHRead was successful at pinpointing important information, as reflected by these assessment parameters. Indeed, *precision* and *recall* were balanced for most of the variables, showing that the EHRead technology is not only accurate when detecting the evaluated variables but also in terms of retrieving a large amount of information.

Although this study deals with EHRs in Spanish, most previous cNLP systems focused on information extraction from clinical reports in English [26]. *F1 scores* of cNLP systems that target EHRs in English range from 0.71 to 0.92 [27-31]. Available rule-based [24,31] or machine learning-oriented [25] systems that identify medical entities in Spanish have reached *F1 scores* between 0.70 and 0.90. However, the cNLP systems targeting the Spanish language are still limited. A direct comparison between the EHRead technology and these state-of-the-art approaches is complicated due to differences in gold standard

creation and use of language. Nevertheless, the overall performance of the EHRead technology across the eight participating sites with the achieved *F1 scores* demonstrates that the performance is comparable to other state-of-the-art NLP systems available in the clinical domain. Furthermore, compared to previous works that detect CD-related variables in English using NLP to increase or correctly classify the number of patients with CD detected through the standard International Classification of Diseases-9 coding system [32,33], our study relies on a purely NLP-dependent detection approach. Having performed our study in Spanish is an added value, since it is a language in which NLP has not been previously applied in CD studies, nonetheless yielding robust results compared to previous approaches in English.

A robust linguistic validation of the EHRead technology sets it forth as a valuable methodology for future studies regarding IBD and CD. The expanding use of EHRs and the wealth of information contained within their free text represent a unique

source of data that benefits from the development of cNLP systems. Indeed, cNLP systems are dynamic and evolve with novel technologies that improve concept identification [21]. This approach is suitable to better detect clinical information of patients with IBD and CD in a real-world setting, which can provide insight to improve the medical management of these patients.

In conclusion, this study presents an evaluation of the EHRead technology, an NLP system for the extraction of clinical information from the narrative free text contained in EHRs. This evaluation clearly demonstrates the ability of the EHRead technology to identify mentions of CD and two related variables. Although further research is needed, the use of the EHRead technology facilitates the automated large-scale analysis of CD, thus contributing to the improvement of clinical practice by generating real-world evidence. Robust data extraction and precise variable detection are key to support future studies using large data sets of patients with CD.

---

## Acknowledgments

We would like to thank Tamara Pozo, Marta Mengual, and Ana Sánchez Gabriel for their kind support during the study, and Stephanie Marchesseau for valuable comments on a previous version of this manuscript. We are grateful to Laura Yebes, Carlos Del Rio-Bermudez, Ana Lopez-Ballesteros, and Clara L Oeste for their assistance in writing and editing the manuscript, and the construction of figures funded by Takeda.

The PREMONITION-CD Study Group includes the following investigators: Carlos Castaño from Hospital Universitario (HU) Rey Juan Carlos, Madrid, Spain; Ángel Ponferrada Díaz from HU Infanta Leonor, Madrid, Spain; María Chaparro and María José Casanova from HU de La Princesa, Madrid, Spain; Felipe Ramos Zabala from HM Hospitales, Madrid, Spain; Almudena Calvache from HU Infanta Elena, Madrid, Spain; Fernando Bermejo from HU de Fuenlabrada, Madrid, Spain; Noemí Manceñido from HU Infanta Sofía, Madrid, Spain; and Marta Calvo Moya from HU Puerta de Hierro, Majadahonda, Madrid, Spain.

This study was funded by Takeda Farmacéutica España S.A. The analyses conducted by Medsavana SL as well as the participation of the Medsavana authors in the development of this manuscript were funded by Takeda Farmacéutica España S.A.

---

## Authors' Contributions

All authors have made substantial contributions to the conception and design of the study, and acquisition, analysis, and interpretation of data, in addition to drafting and revising the manuscript.

---

## Conflicts of Interest

JPG has served as a speaker, a consultant, and advisory member for, or has received research funding from, MSD, Abbvie, Hospira, Pfizer, Kern Pharma, Biogen, Takeda, Janssen, Roche, Sandoz, Celgene, Ferring, Faes Farma, Shire Pharmaceuticals, Dr. Falk Pharma, Tillotts Pharma, Chiesi, Casen Fleet, Gebro Pharma, Otsuka Pharmaceutical, and Vifor Pharma. IG has served as a speaker, a consultant, and advisory member for, or has received research funding from, Kern Pharma, Takeda, and Janssen. RP has served as a speaker for Takeda and Janssen. MIVM has served as a speaker, consultant, and advisory member for, or has received funding from, MSD, Abbvie, Pfizer, Ferring, Shire Pharmaceuticals, Takeda, and Janssen. FG has received educational grants from Janssen, MSD, Takeda, and Abbvie, and nonpersonal investigation grants from MSD, Janssen, Abbvie, Takeda, and Tillotts. CM, JA, VM, IT, and AFN are employees at Takeda Farmacéutica España S.A. LC is an ex-employee and SM is currently employed at Medsavana SL, which received funding from Takeda Farmacéutica España S.A. The remaining authors have no conflicts of interest to declare.

---

## References

1. Freeman HJ. Natural history and long-term clinical course of Crohn's disease. *World J Gastroenterol* 2014 Jan 07;20(1):31-36 [[FREE Full text](#)] [doi: [10.3748/wjg.v20.i1.31](https://doi.org/10.3748/wjg.v20.i1.31)] [Medline: [24415855](https://pubmed.ncbi.nlm.nih.gov/24415855/)]
2. Ananthakrishnan AN. Epidemiology and risk factors for IBD. *Nat Rev Gastroenterol Hepatol* 2015 Apr;12(4):205-217. [doi: [10.1038/nrgastro.2015.34](https://doi.org/10.1038/nrgastro.2015.34)] [Medline: [25732745](https://pubmed.ncbi.nlm.nih.gov/25732745/)]

3. Ramadas AV, Gunesh S, Thomas GAO, Williams GT, Hawthorne AB. Natural history of Crohn's disease in a population-based cohort from Cardiff (1986-2003): a study of changes in medical treatment and surgical resection rates. *Gut* 2010 Sep;59(9):1200-1206. [doi: [10.1136/gut.2009.202101](https://doi.org/10.1136/gut.2009.202101)] [Medline: [20650924](https://pubmed.ncbi.nlm.nih.gov/20650924/)]
4. Del Rio-Bermudez C, Medrano IH, Yebes L, Poveda JL. Towards a symbiotic relationship between big data, artificial intelligence, and hospital pharmacy. *J Pharm Policy Pract* 2020 Nov 09;13(1):75 [FREE Full text] [doi: [10.1186/s40545-020-00276-6](https://doi.org/10.1186/s40545-020-00276-6)] [Medline: [33292570](https://pubmed.ncbi.nlm.nih.gov/33292570/)]
5. Roberts A. Language, structure, and reuse in the electronic health record. *AMA J Ethics* 2017 Mar 01;19(3):281-288 [FREE Full text] [doi: [10.1001/journalofethics.2017.19.3.stas1-1703](https://doi.org/10.1001/journalofethics.2017.19.3.stas1-1703)] [Medline: [28323609](https://pubmed.ncbi.nlm.nih.gov/28323609/)]
6. Sager N, Lyman M, Bucknall C, Nhan N, Tick LJ. Natural language processing and the representation of clinical data. *J Am Med Inform Assoc* 1994;1(2):142-160 [FREE Full text] [doi: [10.1136/jamia.1994.95236145](https://doi.org/10.1136/jamia.1994.95236145)] [Medline: [7719796](https://pubmed.ncbi.nlm.nih.gov/7719796/)]
7. Siddharthan A, Christopher D, Manning and Hinrich Schutze. *Foundations of Statistical Natural Language Processing*. MIT Press, 2000. ISBN 0-262-13360-1. 620 pp. \$64.95/£44.95 (cloth). *Nat Lang Eng* 2002 Jun 17;8(1):91-92. [doi: [10.1017/S1351324902212851](https://doi.org/10.1017/S1351324902212851)]
8. Ford E, Carroll JA, Smith HE, Scott D, Cassell JA. Extracting information from the text of electronic medical records to improve case detection: a systematic review. *J Am Med Inform Assoc* 2016 Sep;23(5):1007-1015 [FREE Full text] [doi: [10.1093/jamia/ocv180](https://doi.org/10.1093/jamia/ocv180)] [Medline: [26911811](https://pubmed.ncbi.nlm.nih.gov/26911811/)]
9. Zheng L, Wang Y, Hao S, Shin AY, Jin B, Ngo AD, et al. Web-based real-time case finding for the population health management of patients with diabetes mellitus: a prospective validation of the natural language processing-based algorithm with statewide electronic medical records. *JMIR Med Inform* 2016 Nov 11;4(4):e37 [FREE Full text] [doi: [10.2196/medinform.6328](https://doi.org/10.2196/medinform.6328)] [Medline: [27836816](https://pubmed.ncbi.nlm.nih.gov/27836816/)]
10. Afzal N, Mallipeddi VP, Sohn S, Liu H, Chaudhry R, Scott CG, et al. Natural language processing of clinical notes for identification of critical limb ischemia. *Int J Med Inform* 2018 Mar;111:83-89 [FREE Full text] [doi: [10.1016/j.ijmedinf.2017.12.024](https://doi.org/10.1016/j.ijmedinf.2017.12.024)] [Medline: [29425639](https://pubmed.ncbi.nlm.nih.gov/29425639/)]
11. Espinosa-Anke L, Tello J, Pardo A, Medrano I, Ureña A, Salcedo I, et al. Savana: a global information extraction and terminology expansion framework in the medical domain. *Procesamiento Lenguaje Nat* 2016;57:23-30.
12. Hernandez Medrano I, Tello Guijarro J, Belda C, Urena A, Salcedo I, Espinosa-Anke L, et al. Savana: re-using electronic health records with artificial intelligence. *Int J Interactive Multimedia Artif Intelligence* 2018;4(7):8. [doi: [10.9781/ijimai.2017.03.001](https://doi.org/10.9781/ijimai.2017.03.001)]
13. Graziani D, Soriano JB, Del Rio-Bermudez C, Morena D, Díaz T, Castillo M, et al. Characteristics and prognosis of COVID-19 in patients with COPD. *J Clin Med* 2020 Oct 12;9(10):3259 [FREE Full text] [doi: [10.3390/jcm9103259](https://doi.org/10.3390/jcm9103259)] [Medline: [33053774](https://pubmed.ncbi.nlm.nih.gov/33053774/)]
14. 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](https://doi.org/10.2196/21801)] [Medline: [33090964](https://pubmed.ncbi.nlm.nih.gov/33090964/)]
15. Izquierdo JL, Almonacid C, González Y, Del Rio-Bermudez C, Ancochea J, Cárdenas R, et al. The impact of COVID-19 on patients with asthma. *Eur Respir J* 2021 Mar;57(3):2003142 [FREE Full text] [doi: [10.1183/13993003.03142-2020](https://doi.org/10.1183/13993003.03142-2020)] [Medline: [33154029](https://pubmed.ncbi.nlm.nih.gov/33154029/)]
16. Jensen PB, Jensen LJ, Brunak S. Mining electronic health records: towards better research applications and clinical care. *Nat Rev Genet* 2012 May 02;13(6):395-405. [doi: [10.1038/mrg3208](https://doi.org/10.1038/mrg3208)] [Medline: [22549152](https://pubmed.ncbi.nlm.nih.gov/22549152/)]
17. 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](https://doi.org/10.1016/j.jbi.2013.06.004)] [Medline: [23810857](https://pubmed.ncbi.nlm.nih.gov/23810857/)]
18. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, et al. Scikit-learn: machine learning in Python. *J Machine Learning Res* 2011;12:2825-2830.
19. McHugh ML. Interrater reliability: the kappa statistic. *Biochem Med (Zagreb)* 2012;22(3):276-282 [FREE Full text] [Medline: [23092060](https://pubmed.ncbi.nlm.nih.gov/23092060/)]
20. Osen H, Chang D, Choo S, Perry H, Hesse A, Abantanga F, et al. Validation of the World Health Organization tool for situational analysis to assess emergency and essential surgical care at district hospitals in Ghana. *World J Surg* 2011 Mar;35(3):500-504 [FREE Full text] [doi: [10.1007/s00268-010-0918-1](https://doi.org/10.1007/s00268-010-0918-1)] [Medline: [21190114](https://pubmed.ncbi.nlm.nih.gov/21190114/)]
21. Canales L, Menke S, Marchesseau S, D'Agostino A, Del Rio-Bermudez C, Taberna M, et al. Assessing the performance of clinical natural language processing systems: development of an evaluation methodology. *JMIR Med Inform* 2021 Jul 23;9(7):e20492 [FREE Full text] [doi: [10.2196/20492](https://doi.org/10.2196/20492)] [Medline: [34297002](https://pubmed.ncbi.nlm.nih.gov/34297002/)]
22. Landis JR, Koch GG. The measurement of observer agreement for categorical data. *Biometrics* 1977 Mar;33(1):159-174. [Medline: [843571](https://pubmed.ncbi.nlm.nih.gov/843571/)]
23. Olivera P, Danese S, Jay N, Natoli G, Peyrin-Biroulet L. Big data in IBD: a look into the future. *Nat Rev Gastroenterol Hepatol* 2019 May;16(5):312-321. [doi: [10.1038/s41575-019-0102-5](https://doi.org/10.1038/s41575-019-0102-5)] [Medline: [30659247](https://pubmed.ncbi.nlm.nih.gov/30659247/)]

24. Oronoz M, Gojenola K, Pérez A, de Ilarraza AD, Casillas A. On the creation of a clinical gold standard corpus in Spanish: mining adverse drug reactions. *J Biomed Inform* 2015 Aug;56:318-332 [[FREE Full text](#)] [doi: [10.1016/j.jbi.2015.06.016](https://doi.org/10.1016/j.jbi.2015.06.016)] [Medline: [26141794](#)]
25. Pérez A, Weegar R, Casillas A, Gojenola K, Oronoz M, Dalianis H. Semi-supervised medical entity recognition: a study on Spanish and Swedish clinical corpora. *J Biomed Inform* 2017 Jul;71:16-30 [[FREE Full text](#)] [doi: [10.1016/j.jbi.2017.05.009](https://doi.org/10.1016/j.jbi.2017.05.009)] [Medline: [28526460](#)]
26. Velupillai S, Mowery D, South BR, Kvist M, Dalianis H. Recent advances in clinical natural language processing in support of semantic analysis. *Yearb Med Inform* 2015 Aug 13;10(1):183-193 [[FREE Full text](#)] [doi: [10.15265/IY-2015-009](https://doi.org/10.15265/IY-2015-009)] [Medline: [26293867](#)]
27. Savova GK, Masanz JJ, Ogren PV, Zheng J, Sohn S, Kipper-Schuler KC, et al. Mayo clinical Text Analysis and Knowledge Extraction System (cTAKES): architecture, component evaluation and applications. *J Am Med Inform Assoc* 2010;17(5):507-513 [[FREE Full text](#)] [doi: [10.1136/jamia.2009.001560](https://doi.org/10.1136/jamia.2009.001560)] [Medline: [20819853](#)]
28. Jonnalagadda SR, Adupa AK, Garg RP, Corona-Cox J, Shah SJ. Text mining of the electronic health record: an information extraction approach for automated identification and subphenotyping of HFpEF patients for clinical trials. *J Cardiovasc Transl Res* 2017 Jun;10(3):313-321. [doi: [10.1007/s12265-017-9752-2](https://doi.org/10.1007/s12265-017-9752-2)] [Medline: [28585184](#)]
29. Liu Z, Yang M, Wang X, Chen Q, Tang B, Wang Z, et al. Entity recognition from clinical texts via recurrent neural network. *BMC Med Inform Decis Mak* 2017 Jul 05;17(Suppl 2):67 [[FREE Full text](#)] [doi: [10.1186/s12911-017-0468-7](https://doi.org/10.1186/s12911-017-0468-7)] [Medline: [28699566](#)]
30. Soysal E, Wang J, Jiang M, Wu Y, Pakhomov S, Liu H, et al. CLAMP - a toolkit for efficiently building customized clinical natural language processing pipelines. *J Am Med Inform Assoc* 2018 Mar 01;25(3):331-336 [[FREE Full text](#)] [doi: [10.1093/jamia/ocx132](https://doi.org/10.1093/jamia/ocx132)] [Medline: [29186491](#)]
31. Moreno I, Moreda P, Romá-Ferri MT. MaNER: A Medical Named Entity Recogniser. Cham: Springer; 2015 Presented at: Natural Language Processing and Information Systems 20th International Conference on Applications of Natural Language to Information Systems; June 17-19, 2015; Passau, Germany p. 418-423. [doi: [10.1007/978-3-319-19581-0\\_40](https://doi.org/10.1007/978-3-319-19581-0_40)]
32. Ananthakrishnan AN, Cai T, Savova G, Cheng S, Chen P, Perez RG, et al. Improving case definition of Crohn's disease and ulcerative colitis in electronic medical records using natural language processing: a novel informatics approach. *Inflamm Bowel Dis* 2013 Jun;19(7):1411-1420 [[FREE Full text](#)] [doi: [10.1097/MIB.0b013e31828133fd](https://doi.org/10.1097/MIB.0b013e31828133fd)] [Medline: [23567779](#)]
33. Kurowski J, Milinovich A, Ji X, Bauman J, Sugano D, Kattan M, et al. Differences in biologic utilization and surgery rates in pediatric and adult Crohn's disease: results from a large electronic medical record-derived cohort. *Inflamm Bowel Dis* 2021 Jun 15;27(7):1035-1044. [doi: [10.1093/ibd/izaa239](https://doi.org/10.1093/ibd/izaa239)] [Medline: [32914165](#)]

## Abbreviations

**CD:** Crohn disease

**cNLP:** clinical natural language processing

**EHR:** electronic health record

**GDPR:** General Data Protection Regulation

**IBD:** inflammatory bowel disease

**NLP:** natural language processing

**SNOMED-CT:** Systemized Nomenclature of Medicine–Clinical Terms

*Edited by C Lovis; submitted 11.05.21; peer-reviewed by D Shung, Y Chen, FJ Sánchez-Laguna; comments to author 29.05.21; revised version received 22.07.21; accepted 02.01.22; published 18.02.22.*

*Please cite as:*

Montoto C, Gisbert JP, Guerra I, Plaza R, Pajares Villarroya R, Moreno Almazán L, López Martín MDC, Domínguez Antonaya M, Vera Mendoza I, Aparicio J, Martínez V, Tagarro I, Fernandez-Nistal A, Canales L, Menke S, Gomollón F, PREMONITION-CD Study Group

*Evaluation of Natural Language Processing for the Identification of Crohn Disease–Related Variables in Spanish Electronic Health Records: A Validation Study for the PREMONITION-CD Project*

*JMIR Med Inform* 2022;10(2):e30345

URL: <https://medinform.jmir.org/2022/2/e30345>

doi: [10.2196/30345](https://doi.org/10.2196/30345)

PMID: [35179507](https://pubmed.ncbi.nlm.nih.gov/35179507/)

©Carmen Montoto, Javier P Gisbert, Iván Guerra, Rocío Plaza, Ramón Pajares Villarroya, Luis Moreno Almazán, María Del Carmen López Martín, Mercedes Domínguez Antonaya, Isabel Vera Mendoza, Jesús Aparicio, Vicente Martínez, Ignacio Tagarro, Alonso Fernandez-Nistal, Lea Canales, Sebastian Menke, Fernando Gomollón, PREMONITION-CD Study Group. Originally



published in JMIR Medical Informatics (<https://medinform.jmir.org>), 18.02.2022. This is an open-access article distributed under the terms of the Creative Commons Attribution License (<https://creativecommons.org/licenses/by/4.0/>), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Medical Informatics, is properly cited. The complete bibliographic information, a link to the original publication on <https://medinform.jmir.org/>, as well as this copyright and license information must be included.

Original Paper

# Traditional Machine Learning Models and Bidirectional Encoder Representations From Transformer (BERT)–Based Automatic Classification of Tweets About Eating Disorders: Algorithm Development and Validation Study

José Alberto Benítez-Andrades<sup>1\*</sup>, BSc, MSc, PhD; José-Manuel Alija-Pérez<sup>2</sup>, BSc, MSc, PhD; Maria-Esther Vidal<sup>3\*</sup>, BSc, MSc, PhD; Rafael Pastor-Vargas<sup>4\*</sup>, BSc, MSc, PhD; María Teresa García-Ordás<sup>2\*</sup>, BSc, MSc, PhD

<sup>1</sup>SALBIS Research Group, Department of Electric, Systems and Automatics Engineering, University of León, León, Spain

<sup>2</sup>SECOMUCI Research Group, Escuela de Ingenierías Industrial e Informática, Universidad de León, León, Spain

<sup>3</sup>Leibniz University of Hannover, Hannover, Germany

<sup>4</sup>Communications and Control Systems Department, Spanish National University for Distance Education, Madrid, Spain

\*these authors contributed equally

**Corresponding Author:**

José Alberto Benítez-Andrades, BSc, MSc, PhD  
SALBIS Research Group  
Department of Electric, Systems and Automatics Engineering  
University of León  
Campus of Vegazana s/n  
León, 24071  
Spain  
Phone: 34 987293628  
Email: [jbena@unileon.es](mailto:jbena@unileon.es)

## Abstract

**Background:** Eating disorders affect an increasing number of people. Social networks provide information that can help.

**Objective:** We aimed to find machine learning models capable of efficiently categorizing tweets about eating disorders domain.

**Methods:** We collected tweets related to eating disorders, for 3 consecutive months. After preprocessing, a subset of 2000 tweets was labeled: (1) messages written by people suffering from eating disorders or not, (2) messages promoting suffering from eating disorders or not, (3) informative messages or not, and (4) scientific or nonscientific messages. Traditional machine learning and deep learning models were used to classify tweets. We evaluated accuracy, F1 score, and computational time for each model.

**Results:** A total of 1,058,957 tweets related to eating disorders were collected. were obtained in the 4 categorizations, with The bidirectional encoder representations from transformer–based models had the best score among the machine learning and deep learning techniques applied to the 4 categorization tasks (F1 scores 71.1%-86.4%).

**Conclusions:** Bidirectional encoder representations from transformer–based models have better performance, although their computational cost is significantly higher than those of traditional techniques, in classifying eating disorder–related tweets.

(*JMIR Med Inform* 2022;10(2):e34492) doi:[10.2196/34492](https://doi.org/10.2196/34492)

**KEYWORDS**

natural language processing; NLP; social media; data; bidirectional encoder representations from transformer; BERT; deep learning; machine learning; eating disorder; mental health; model; classification; Twitter; nutrition; diet; weight; disorder; performance

## Introduction

### Background

Physical appearance is an essential element for people in this society. Although many studies corroborate that moderate physical activity and proper nutrition help to maintain a healthy body [1] and mind [2], a large part of society continues to place more importance on physical appearance than on health. In recent years, trends have promoted a curvy physique [3,4] despite it being unhealthy, and most people associate having a slim body with being happy to have a slim body. These associations between physical appearance and happiness are the causes of illnesses such as eating disorders. These mental illnesses are complex and do not depend on a single factor [5,6]. Thus, messages relating being fat or skinny with aesthetics that are contained in some media—advertisements, magazines, and celebrity social media—can hurt people vulnerable to these types of illnesses.

The prevalence of eating disorders has been increasing [7]. In addition, since the start of the COVID-19 pandemic, there has been a more pronounced increase in eating disorders [7]. Therefore, any strategy that helps to combat this health problem may be of interest to society.

With the emergence of social media, studies [8-12] using social media data to propose solutions that can help combat this type of illness from different perspectives have also emerged. Artificial intelligence and machine learning techniques, mainly applied to text, have been used to find patterns that help in classifying text to explore eating disorder-related discourse shared on Twitter [8,9] and other studies [10-12] are making use of the data contained in social networks and offering solutions that can help in the field of public health. Among the social media currently available, the most widely used platform in scientific studies is Twitter [13].

Despite the increase in studies on eating disorders that have, for example, analyzed pro-eating disorder websites [14], performed sentiment analysis of pro-anorexia and anti-proanorexia videos on YouTube [15], and that have used social media data and artificial intelligence techniques on pro-eating disorder and prorecovery text [16], none has identified (1) tweets that have been written by people who suffer or have suffered from eating disorders, (2) tweets that promote eating disorders, (3) informative or noninformative tweets related to eating disorders and, within the informative tweets, (4) which ones make use of scientific information and which ones do not.

### Objectives

Our main objectives were to achieve accurate text classification in performing these 4 tasks, to compare the efficiency of text classification models using traditional machine learning techniques and those using novel techniques, such as pretrained bidirectional encoder representations from transformer (BERT)-based models, to determine which approach has the best combination of performance and computational cost and would be useful for future research.

In our previous research [17], in which 6 test-beds were conducted, the main objective was only to apply 6 pretrained bidirectional encoder representations from transformer-based models to classify a category in a data set. This time, we used a broader approach, by presenting the main problem as a comparison of the performance (accuracy and computational cost) of traditional machine learning models vs bidirectional encoder representations from transformer-based models on 4 different data categorization tasks. This meant increasing 6 test-beds to 40 different test-beds.

### Literature Review

#### *Social Media in Health Informatics*

Social media, specifically social networks, have become very important sources of information within the field of health informatics. Health informatics includes the design and application of innovations based on information technologies to solve problems related to public health and health services [18]. In this branch of interdisciplinary science, it is possible to carry out complex research to manage information to improve efficiency and reduce costs in health care [19]. Health informatics includes information science, informatics, and health care.

Health-related research using social media is mainly focused on two areas. In real-time monitoring and the prediction of diseases (eg, influenza), it is possible to collect and use messages that have been geographically localized and that are on topics of interest. In this way, research tasks related to the user discussions are a simple task. Social media are also used to determine perspectives on different health problems and conditions. Thus, social media are useful, easy to use, and very important tools for observational studies.

Twitter is a very popular and widely used social network within the field of health and social health research. Some studies [9,14-16] make use of artificial intelligence techniques, such as social network data mining, to generate predictive models based on current knowledge. These techniques have been used, for example, in the context of the COVID-19 pandemic, to determine the public's perceptions [20] and to examine communication behavior between health organizations and users [21].

Studies [8,9,17,22-24] in the field of health informatics have used Twitter to study user behaviors and characteristics such as location, frequency, most used hashtags, or the structure of user networks. This information, being public and anonymous, is typically exempt from requiring the approval of an ethics committee [25]. Other studies have analyzed the impact of content shared by users [21] and how Twitter is used to receive and provide emotional support [26] or to determine the best way raise public awareness (World Rare Disease Day [27]).

Social media facilitate a great deal of research in the field of health informatics, for example, sentiment analysis, behavioral analysis, or information dissemination analysis, which make use of techniques related to machine learning or deep learning techniques for the classification and prediction of content that has been prepared using natural language processing.

### **Classification Methods and Health Informatics**

Supervised machine learning techniques are used to predict an outcome based on a given input by constructing an input–output pair. The main goal is to build a model that can then be used to make accurate predictions using new data.

Tasks in the field of supervised machine learning include regression—the prediction of a real number—and classification—the prediction of a class label [28]. Supervised classification tasks make use of a labeled training data set. This set allows the creation of classifiers or predictive models [28]. Text mining techniques are used to quantify text data (what is feature engineering) to represent the relationships between words as tokens.

Classification techniques make it possible to categorize large data sets efficiently to study text-based data. This approach has many advantages—more accurate predictions than those of humans and time savings [29-31]. Some commonly used classification techniques in health informatics are logistic regression, support vector machines, Naïve Bayes, random forest, gradient boosting trees, decision trees, and gradient boosted regression trees.

Naïve Bayes classifiers have been used to predict Zika and dengue diseases using data obtained from Twitter [32] and to test the classification of 4 conditions—influenza, depression, pregnancy, and eating disorders—and 2 locations—Portugal and Spain [33].

Other studies [34-36] have shown that good results can be obtained using support vector machine algorithms, such as, with a neural network to predict COVID-19 in chest x-ray images, a prediction model [35], and for sentiment analysis tasks on a Twitter data set related to the COVID-19 pandemic in Canada [36]. A gradient boosted regression tree classifier was used to identify tweets related to e-cigarettes [22], with accurate classification of 5 different user types, by manually labeling a sample of tweets and using feature engineering techniques based on the term frequency–inverse document frequency matrix.

It is also possible to combine different classification algorithms and compare their performance to use the best performing classifier for a given task [37], for example, gradient boosting tree, decision tree, logistic regression, and support vector machine models were used to predict patient needs at the level of informational support [23].

### **Social Media Research Related to Eating Disorders**

There are a number of studies that make use of data related to eating disorders [8,14-16,24,38]. In one study [24], 123,977 tweets were collected and a subsample of 2219 was labeled; the

efficiency of a convolutional neural network, with long short-term memory, in classifying tweets about eating disorders was demonstrated. Another study [8] statistically analyzed the effect of eating disorder awareness campaigns by obtaining information on tweets that mentioned 2 hashtags. A review [39] showed the importance of machine learning in advancing the prediction, prevention, and treatment of mental illness and eating disorders. Other studies [14-16,38] have demonstrated the importance of the use of data obtained from social networks in the field of eating disorders, by performing analysis from a social rather than computational perspective, which is known as social network analysis.

A previous social media study predicted depression from texts [40]; therefore, detecting texts written by people suffering from eating disorders can also be helpful. Studies on the detection of pro-ana and pro-recovery communities [41-43]—people in favor of and who promote anorexia and recovery from eating disorders, respectively—and reviews [44,45], have suggested this type of study may be useful. Furthermore, to the best of our knowledge, no studies having the same objectives as ours have been conducted.

## **Methods**

### **Data Collection**

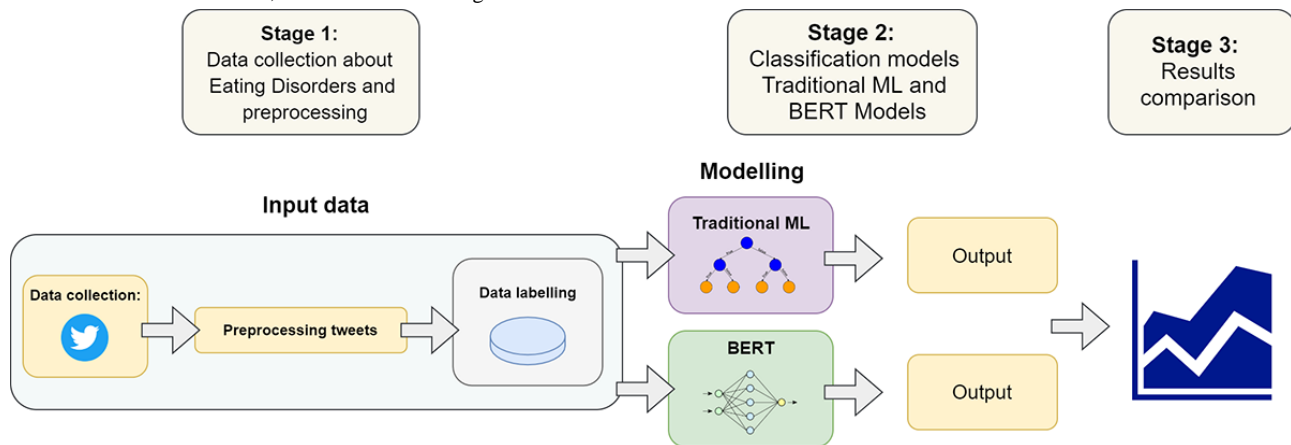
#### **Tweets**

A tool (T-Hoarder [46]) was used to collect tweets (Figure 1). Tweets were obtained at the moment they were sent because the tool uses the Twitter streaming API, thus tweets that were subsequently removed from the platform for not complying with regulations were still obtained.

T-Hoarder allowed us to obtain additional information about tweets for further analysis, such as, ID, text, and author (among other fields). Tweets were identified by keywords [17]. In set 1 “anorexia,” “anorexic,” “dietary disorders,” “inappetence,” “feeding disorder,” “food problem,” “binge eating,” and “anorectic” we used. In set 2, “eating disorders,” “bulimia,” “food issues,” “loss of appetite,” “food issue,” “food hater,” “eat healthier,” “disturbed eating habits,” “abnormal eating habits,” and “abnormal eating habit” were used. In set 3, “binge-vomit syndrome,” “bingeing,” “bulimarexia,” “anorexic skinny,” and “eating healthy” were used.

By using a different Twitter accounts for each set, more tweets could be obtained without exceeding the Twitter platform's usage limit. English terms were used because more tweets are generated in English [46].

**Figure 1.** Study workflow: (A) data collection and preprocessing, (B) classification model training, and (C) evaluation. BERT: bidirectional encoder representations from transformer, ML: machine learning.



### Preprocessing

Preprocessing was conducted in Python (version 3.6). Data were loaded from documents obtained through T-Hoarder, which generates a file up to 100 MB; therefore, 4 files were obtained for data set 1, 4 files were obtained for data set 2, and 2 files were obtained for data set 3. Some data, such as location, name, and biography, contained line breaks or tabs. To avoid conflicts with delimiters, tabs and line breaks were removed using a function. After preprocessing the data frames, they were concatenated into a single data frame. In order to be able to work in a more agile way with the data frame, the memory usage of the data frame was calculated and optimized by converting numeric columns into numbers, converting dates to datetime format, and converting the remaining objects into categories. These steps helped reduce the data frame from 2.7 GB to 1.1 GB. We removed all tweets that were retweets, duplicates (because we unified data sets that might contain common tweets), and non-English tweets.

To select the subset of 2000 tweets, manual filtering was performed to eliminate tweets that were not related to eating disorder issues. Some of our keywords were too generic and meant that the tweets collected were not about eating disorders. For example, some of these words that triggered the collection of tweets unrelated to eating disorders were “food problem,” “inappetence,” “food issue,” and “bingeing”; however, in order to generate predictive models with greater accuracy and less bias, we kept a small sample of tweets ( $n=286$ ) that did not belong to any of the categories, but that did contain some of the keywords of interest.

### Labeling

Tweets in 4 different categories in the subset were manually labeled (Table 1). Labeling was carried out by 2 people, labeling 1000 tweets each. The labels were then reviewed by 4 mental illness experts. This procedure took place over the course of 1 full month, with each person taking approximately a total of 70 hours in carrying out this work. In category 1, tweets written by people suffering from eating disorders were represented with a value of 1, and the rest were represented with a value of 0. To assess this, each user profile was accessed and user description and tweets published by the user were examined to determine if the user had publicly mentioned having an eating disorder. In category 2, tweets that promoted having an eating disorder were labeled with a value of 1, and all other tweets were labeled with a value of 0. There are communities of people who suffer from eating disorders who try to encourage other people to also suffer from it by promoting it as if it were something positive or fashionable. There are many studies [9] that talk about pro-eating disorders communities using the terms *pro-ana* or *pro-anorexia*. In category 3, informative tweets were represented with a value of 1, and noninformative tweets were represented with a value of 0. Informative tweets are those that show information with the aim of informing readers, while the rest were those in which the author expressed an opinion. In category 4, scientific tweets were labeled with a value of 1, and the rest were labeled with a value of 0. A tweet of an informative nature that had been written by a person belonging to the field of research, for example, a doctor of philosophy in different subjects, was labeled as a scientific tweet. Scientific tweets were also those that shared links to papers published in scientific journals. If a tweet did not belong to any of the 4 proposed categories, it was not eliminated from the data set, since having tweets with value of 0 was also necessary.

**Table 1.** Categories of labeled tweets and examples.

Category topics	Tweet
<b>Category 1</b>	
Written by someone who suffers from eating disorder	i was stressed and ate a whole bowl of pasta, where's my badge for being the worst anorexic #edwtw
Written by someone who does not have an eating disorder	Is your #teenager not eating or eating a lot less than normal? She might be suffering from #anorexia. We can help; please come see us <a href="https://t.co/GfStM1IVGz">https://t.co/GfStM1IVGz</a> #weightloss #losingweight <a href="https://t.co/z5NK0tjNIi">https://t.co/z5NK0tjNIi</a>
<b>Category 2</b>	
Promotes eating disorders	Currently feeling like the best anorexic #eating disordertwt <a href="https://t.co/1BZPMs8bGU">https://t.co/1BZPMs8bGU</a> #mentalhealth #diet #anorexia
Not promotes eating disorders	Higher-calorie diets could lead to a speedier recovery in patients with anorexia nervosa, study shows <a href="https://t.co/mipX3nrhHN">https://t.co/mipX3nrhHN</a>
<b>Category 3</b>	
Informative	#AnorexiaNervosa – A Father and Daughter Perspective -Highlights from RCPsychIC 2019 # EatingDisorders #mentalhealth <a href="https://t.co/iq3GH5ce6C">https://t.co/iq3GH5ce6C</a>
Noninformative	Binge eating makes me sad :( #eatingdisorder #bingeeating <a href="https://t.co/0jff7YrVyc">https://t.co/0jff7YrVyc</a>
<b>Category 4</b>	
Scientific	The problem extends to Food and Drug Administration and National Institutes of Health data sets used in a recent study appearing in Reproductive Toxicology. #ai #technology #BigData #ML <a href="https://t.co/DFvh6gNA38">https://t.co/DFvh6gNA38</a>
Nonscientific	Do not waste time thinking about what you could have done differently. Keep your eyes on the road ahead and do it differently now. #anorexia #eatingdisorder #recovery #nevergiveup #alwayskeepfighting <a href="https://t.co/YalYzclBDM">https://t.co/YalYzclBDM</a>

### Final Sample

Before training and validating the models, tweets in the labeled set with more than 80% similarity were eliminated. It was decided to apply this criterion for tweets containing the same text but using different hashtags. Remaining tweets were processed by removing the stop words (words that have no meaning on their own and that modify or accompany other words, for example, articles, pronouns, adverbs, prepositions, or some verbs) and punctuation or symbols, that hindered the application of machine learning techniques.

### Classification Methods

#### General

We used random forest, recurrent neural networks, bidirectional long short-term memory networks (ClassificationModel; simpletransformers [47], version 0.62.2), and pretrained bidirectional encoder representations from transformer-based models (RoBERTa [48], BERT [49], CamemBERT [50], DistilBERT [51], FlauBERT [52], ALBERT [53], and RobBERT [54]). Bidirectional long short-term memory and bidirectional encoder representations from transformer-based methods were chosen because they seemed to be the most promising models for natural language processing [55-57]. In addition, random forest was used for comparison because it is a traditional machine learning technique.

Two models—CamemBERT [50] and FlauBERT [52]—were pretrained using French text, and RobBERT [54] was pretrained using Dutch text. We used these models to obtain performance

data for with text not written in their initial language. Data were divided into 70% training and 30% testing sets (train\_test\_split function in scikit-learn). The evaluation metrics were accuracy and F1 score.

For the random forest model, 5-fold cross-validation was used. For the neural networks, 5 different iterations were performed, and the mean F1 score and accuracy were obtained.

#### Random Forest

Random forest models [58] are constructed from a set of decision trees, which are usually trained with a method called bagging, to take advantage of the independence between the simple algorithms, since error can be greatly reduced by averaging the outputs of the simple models. Several decision trees are built and fused in order to obtain a more stable and accurate prediction. Random forest models can be used for both regression and classification problems.

One of the advantages offered by this type of model is the additional randomness when more trees are included. The algorithm searches for the best feature as a node is split from a random set of features. This makes it possible to obtain models with better performance. When a node is split, only a random subset of features is considered. Random thresholds can also be used for each feature, instead of searching for the best possible threshold, which adds additional randomness.

#### Recurrent Neural Network

In this type of neural network, a temporal sequence that contains a directed graph made up of connections between different nodes

is defined. These networks have the capacity to show a dynamic temporal behavior. These types of networks, which are derived from feedforward neural networks, have the ability to use memory (their internal state) to process input sequences of varying lengths. This feature makes recurrent neural networks useful for tasks such as unsegmented and connected handwriting recognition or speech recognition [55,59,60].

There are 2 classes of recurrent neural networks—finite-pulse and infinite-pulse. The former are made up of a directed acyclic graph that can be unrolled and replaced by a strictly feedforward neural network, whereas the latter are made of a directed cyclic graph, which does not allow the graph from being unrolled.

### ***Bidirectional Long Short-Term Memory***

Bidirectional long short-term memory networks [61] are constructed from 2 long short-term memory modules that, at each time step, take past and future states into account to produce the output.

### ***Bidirectional Encoder Representations From Transformer-Based Models***

The bidirectional encoder representations from transformer framework is not a model in itself. According to Devlin et al [49], it is a “language understanding” model.

In the bidirectional encoder representations from transformer-based method, a neural network is trained to learn a language, similar to transfer learning in computer vision neural networks, and follows the linguistic representation in a bidirectional way, looking at the words both after and before

each words. It is the combination of these approaches that has made it a successful natural language processing method [62].

### ***Configuration***

We used Jupyter notebook and TensorFlow and Pytorch libraries. It was necessary to use both libraries because, currently, bidirectional encoder representations from transformer-based networks can only be generated through Pytorch, while TensorFlow is one of the most widely used libraries to generate random forest, recurrent neural network, and bidirectional long short-term memory models.

### ***Hyperparameters***

We used a grid search (GridSearchCV) to select the random forest parameters (Table 2).

To train recurrent neural networks (sklearn; keras) to perform the binary categorization tasks, the sigmoid activation function used (Figure 2). We trained and validated the bidirectional long short-term memory models (sklearn and TensorFlow libraries) using the best-performing configuration (Figure 3), after carrying out different tests.

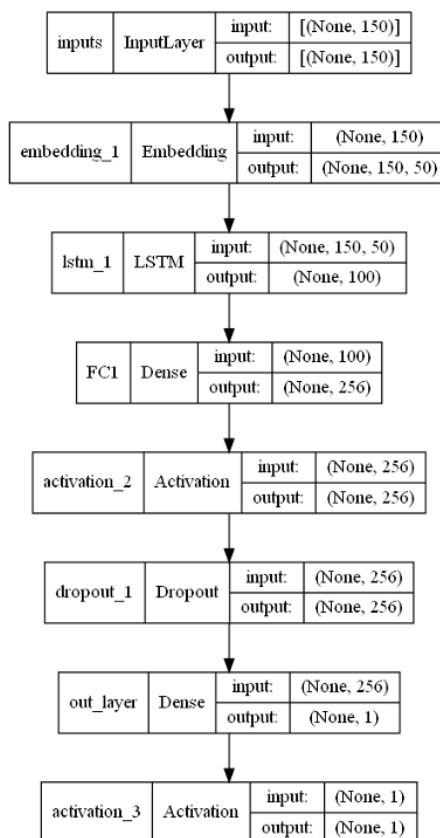
For the 7 pretrained bidirectional encoder representations from transformer-based models, the hyperparameters were `reprocess_input_data=True; fp16=False; evaluate_during_training=False; evaluate_during_training_verbose=False; learning_rate=2e-5; train_batch_size=32; eval_batch_size=32; num_train_epochs=15; overwrite_output_dir=True; and evaluation_strategy='epochs'`.

All experiments and data are published in a repository accessible to anyone [63].

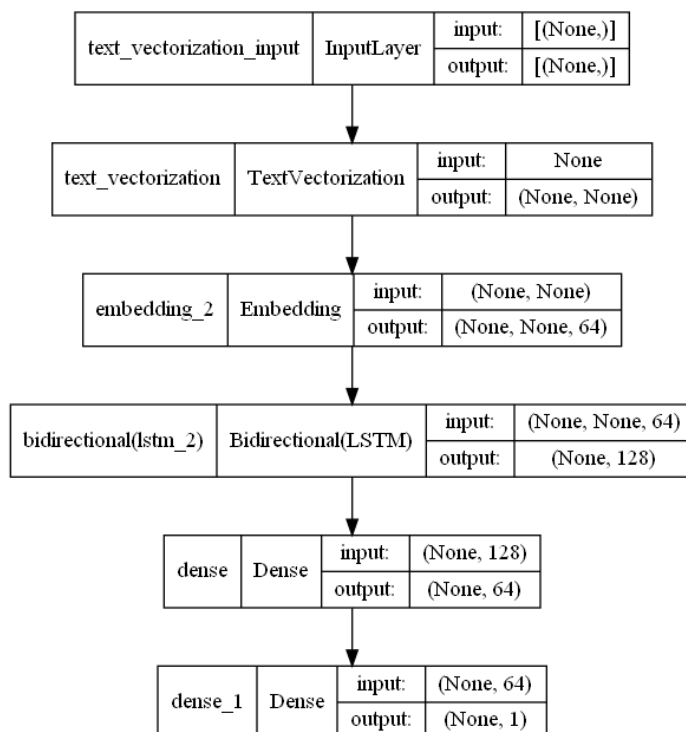
**Table 2.** Random forest hyperparameters.

Category	criterion	max_depth	max_features	n_estimators
Category 1	gini	7	log2	200
Category 2	gini	8	auto	1000
Category 3	gini	8	sqrt	800
Category 4	gini	8	auto	1000

**Figure 2.** Architecture of the recurrent neural network. LSTM: long short-term memory.



**Figure 3.** Architecture of the bidirectional long short-term memory (LSTM) network.



## Results

### Preprocessing

A total of 1,085,957 tweets, written and posted on Twitter between October 20, 2020 and December 26, 2020, were

collected. After preprocessing, a total of 494,025 valid tweets were obtained. These tweets are shared and publicly available on the Kaggle platform [64]. From the subset of 2000 tweets that was manually labeled, 1877 remained after the similarity criterion was applied. Table 3 shows the 10 most repeated terms in the full set of tweets and in the subset that was labeled.



**Table 3.** Table of terms and frequencies of the 10 most repeated terms in the initial data set and in the labeled subset of data.

Term	Frequency, n
<b>Complete set (n=494,025)</b>	
hey mp	230,013
healthy	210,430
pltpinkmonday	209,330
eat	183,436
covid19	156,541
edtw	123,175
anorexia	112,864
disorders	102,063
endsars	99,844
bachelorette	48,370
problem	45,959
<b>Subset (n=2000)</b>	
eat	1132
disorder	830
food	410
recovery	382
edtw	301
binge	282
people	245
anorexic	244
research	226
study	202
problem	199

## Category

In category 1, 50.2% (942/1877) of tweets were written by a person with an eating disorder, and 49.8% (935/1877) of tweets were written by a person without an eating disorder. In category 2, 23.8% (447/1877) of tweets encourage people to suffer from an eating disorder, and 76.2% (1400/1877) of tweets do not encourage people to suffer from an eating disorder.

In category 3, 37% (694/1877) of the tweets were informative, 63% (1183/1877) of tweets were opinionated. In category 4, 23.3% (437/1877) of the tweets were scientific, 76.7% (1440/1877) of tweets were of a nonscientific nature.

## Performance

Performance (Table 4) and implementation time (Table 5), which corresponds to the time invested in generating and validating the different models, for 4 different categorization tasks. The pretrained RoBERTa model was the most accurate

for detecting tweets that had been written by people suffering from some type of eating disorder (accuracy 83.1%). Despite this, the more traditional recurrent neural network yielded an accuracy that was not much lower (accuracy 82.6%). The most accurate model for the detection of tweets that did or did not promote an eating disorder was the RoBERTa model (accuracy 88.5%); however, applying bidirectional long short-term memory improved performance (accuracy 86.7%). The most accurate model for the detection of informative or opinion-based tweets was RoBERTa model (accuracy 84.4%). Accuracy for all bidirectional encoder representations from transformer-based models, except ALBERT and FlauBERT, exceeded 80%; however, applying bidirectional long short-term memory resulted in an accuracy of 78.7%. The model with the highest accuracy for the detection of scientific or nonscientific tweets was the RoBERTa model (accuracy 94.2%). All bidirectional encoder representations from transformer-based models equaled or exceeded 92%; however, applying bidirectional long short-term memory yielded an accuracy of 85.8%.

**Table 4.** Classification performance.

Model	Having eating disorders or not		Encouraging eating disorders or not		Informative or not		Scientific or not	
	F1 score, %	Accuracy, %	F1 score, %	Accuracy, %	F1 score, %	Accuracy, %	F1 score, %	Accuracy, %
Random forest	79.8	79.2	47	76.7	49.2	73.7	27.3	80.4
Recurrent neural network	83.2	82.6	61	82.1	67.3	70.7	67.3	70.7
Bidirectional long short-term memory	78.5	79.3	67.1	86.7	67.1	78.7	76.8	85.8
Bidirectional encoder representations from transformer-based <sup>a</sup>	83.3	83	71.9	87.2	77.6	84.3	86	94.1
RoBERTa <sup>a</sup>	83.8	83.1	74.3	88.5	77.6	84.4	86.4	94.2
DistilBERT <sup>a</sup>	84	83.1	72.3	87.3	75	82.8	84.2	93.3
CamemBERT <sup>a</sup>	79.1	78.7	73.6	87.8	74.7	81.7	82.5	92.3
ALBERT <sup>a</sup>	81.2	80.4	74.3	88.2	73.8	81.5	83.3	93
FlauBERT <sup>a</sup>	82.6	81.7	72.9	87.5	72.2	80	83.4	92.7
RobBERT <sup>a</sup>	78.8	78.4	71.1	86.2	73.8	81.6	83	92.6

<sup>a</sup>A pretrained model was used: bert-based-multilingual-cased for BERT, roberta-base for RoBERTa, distilbert-base-cased for DistilBERT, camembert-base for CamemBERT, albert-base-v1 for ALBERT, flaubert-base-cased for FlauBERT, and robert-v2-dutch-base for RobBERT.

For bidirectional encoder representations from transformer-based models, despite obtaining better performance metrics in terms of accuracy, the training and validation times of the models are much higher than those of random forest, recurrent neural network, and bidirectional long short-term memory models. For example, bidirectional encoder representations from transformer-based models take approximately 15 times longer than random forest models (Table 5).

The improvements between the accuracy of the best bidirectional encoder representations from transformer-based model

(Categorization 1: DistilBERT 83.1%; Categorization 2: RoBERTa 88.5%; Categorization 3: RoBERTa 84.4%; Categorization 4: RoBERTa 94.2%) and that of the best model between random forest, recurrent neural network, or bidirectional long short-term memory models (Categorization 1: recurrent neural network 82.6%; Categorization 2: bidirectional long short-term memory 86.7%; Categorization 3: bidirectional long short-term memory 78.7%; Categorization 4: bidirectional long short-term memory 85.8%) were 0.61%, 2.08%, 7.24%, and 9.79%, respectively.

**Table 5.** Implementation time.

Model	Time (seconds)			
	Having eating disorders or not	Encouraging eating disorders or not	Informative or not	Scientific or not
Random forest	1.74	12.8	10.4	12.9
Recurrent neural network	152.1	163.1	151.5	153.7
Bidirectional long short-term memory	163.2	175.3	164.8	167.9
Bidirectional encoder representations from transformer-based	1257.4	1232.1	1292.7	1311.4
RoBERTa	1116.2	1158.8	1142.5	1192.8
DistilBERT	1343.3	1327.8	1332.0	1362.3
CamemBERT	1472.3	1457.5	1462.0	1493.4
ALBERT	1372.7	1352.3	1331.3	1392.5
FlauBERT	1203.9	1207.1	1202.1	1235.1
RobBERT	1234.4	1215.4	1319.7	1123.5

## Discussion

### Principal Results

Practitioners and researchers can benefit from the use of social media data in the field of eating disorder. Although the model with the best accuracy was always one of the pretrained bidirectional encoder representations from transformer-based models, the computational costs compared with those of simpler models may be excessive. The difference between the accuracy of the best bidirectional encoder representations from transformer-based model and the best of the 3 simpler models (random forest, recurrent neural network, and bidirectional long short-term memory) did not exceed 9.79%.

Given the high computational cost, use of bidirectional encoder representations from transformer-based models in this instance may not be essential. The accuracy for the 4 different categorization tasks is relatively high even in the simplest models.

Despite the fact that we used only 1877 tweets (which is similar to the amounts used in previous studies: 2219 [24] and 2095 [65]), the models classified the tweets with a high level of accuracy.

For the classification of tweets into informative or noninformative (categorization 3), our models obtained a higher accuracy (80%-84.4%) than those in previous studies (77.7% [44] and 81% [45]). Comparisons cannot be made for the other 3 categorization tasks because of a lack of applicable eating disorder-related studies.

### Limitations

This research has several limitations. (1) It was limited to a social media platform, (2) some categorization tasks were not balanced, which may lead to bias in the generated models, (3) the training set was sufficient but could be larger for better results in a real environment, and (4) when labeling tweets, it is possible that there was a bias in determining whether a tweet was written by someone with an eating disorder due to lack of information about the user.

### Conclusions

Machine learning and deep learning models were used to classify eating disorder-related tweets into binary categories in 4 categorization tasks, with accuracies greater than 80%. The best performing models were RoBERTa and DistilBERT, both bidirectional encoder representations from transformer-based classification methods.

The computational cost was much higher for the bidirectional encoder representations from transformer-based models compared to those of the simpler models (random forest, recurrent neural network or traditional bidirectional long short-term memory), time invested in training and validation was greater by a factor of 10.

Future work will include (1) increasing the training and validation data set, (2) applying natural language processing techniques that make use of ontologies with which it is possible to include automation and logical reasoning, (3) integrating predictive models in a real-world development project, such as a Twitter bot, and (4) validating the model using texts written by patients with eating disorders and who are in treatment.

### Conflicts of Interest

None declared.

### References

1. Norris E, van Steen T, Direito A, Stamatakis E. Physically active lessons in schools and their impact on physical activity, educational, health and cognition outcomes: a systematic review and meta-analysis. *Br J Sports Med* 2020 Jul 16;54(14):826-838. [doi: [10.1136/bjsports-2018-100502](https://doi.org/10.1136/bjsports-2018-100502)] [Medline: [31619381](https://pubmed.ncbi.nlm.nih.gov/31619381/)]
2. Thomas J, Thirlaway K, Bowes N, Meyers R. Effects of combining physical activity with psychotherapy on mental health and well-being: a systematic review. *J Affect Disord* 2020 Mar 15;265:475-485. [doi: [10.1016/j.jad.2020.01.070](https://doi.org/10.1016/j.jad.2020.01.070)] [Medline: [32090775](https://pubmed.ncbi.nlm.nih.gov/32090775/)]
3. Izquierdo A, Plessow F, Becker KR, Mancuso CJ, Slattery M, Murray HB, et al. Implicit attitudes toward dieting and thinness distinguish fat-phobic and non-fat-phobic anorexia nervosa from avoidant/restrictive food intake disorder in adolescents. *Int J Eat Disord* 2019 Apr 31;52(4):419-427 [FREE Full text] [doi: [10.1002/eat.22981](https://doi.org/10.1002/eat.22981)] [Medline: [30597579](https://pubmed.ncbi.nlm.nih.gov/30597579/)]
4. Urdapilleta I, Lahlou S, Demarchi S, Catheline J. Women with obesity are not as curvy as they think: consequences on their everyday life behavior. *Front Psychol* 2019 Aug 16;10:1854 [FREE Full text] [doi: [10.3389/fpsyg.2019.01854](https://doi.org/10.3389/fpsyg.2019.01854)] [Medline: [31474907](https://pubmed.ncbi.nlm.nih.gov/31474907/)]
5. Griffen TC, Naumann E, Hildebrandt T. Mirror exposure therapy for body image disturbances and eating disorders: a review. *Clin Psychol Rev* 2018 Nov;65:163-174. [doi: [10.1016/j.cpr.2018.08.006](https://doi.org/10.1016/j.cpr.2018.08.006)] [Medline: [30223161](https://pubmed.ncbi.nlm.nih.gov/30223161/)]
6. Hoek H. Review of the worldwide epidemiology of eating disorders. *Curr Opin Psychiatry* 2016 Nov;29(6):336-339. [doi: [10.1097/YCO.0000000000000282](https://doi.org/10.1097/YCO.0000000000000282)] [Medline: [27608181](https://pubmed.ncbi.nlm.nih.gov/27608181/)]
7. Qian J, Wu Y, Liu F, Zhu Y, Jin H, Zhang H, et al. An update on the prevalence of eating disorders in the general population: a systematic review and meta-analysis. *Eat Weight Disord* 2021 Apr 08. [doi: [10.1007/s40519-021-01162-z](https://doi.org/10.1007/s40519-021-01162-z)] [Medline: [33834377](https://pubmed.ncbi.nlm.nih.gov/33834377/)]

8. Viguria I, Alvarez-Mon MA, Llaverro-Valero M, Asunsolo Del Barco A, Ortuño F, Alvarez-Mon M. Eating disorder awareness campaigns: thematic and quantitative analysis using Twitter. *J Med Internet Res* 2020 Jul 14;22(7):e17626 [FREE Full text] [doi: [10.2196/17626](https://doi.org/10.2196/17626)] [Medline: [32673225](https://pubmed.ncbi.nlm.nih.gov/32673225/)]
9. Sukunesan S, Huynh M, Sharp G. Examining the pro-eating disorders community on Twitter via the hashtag #proana: statistical modeling approach. *JMIR Ment Health* 2021 Jul 09;8(7):e24340 [FREE Full text] [doi: [10.2196/24340](https://doi.org/10.2196/24340)] [Medline: [34255707](https://pubmed.ncbi.nlm.nih.gov/34255707/)]
10. Fiumara G, Celesti A, Galletta A, Carnevale L, Villari M. Applying artificial intelligence in healthcare social networks to identify critical issues in patients' posts. In: *Proceedings of the 11th International Joint Conference on Biomedical Engineering Systems and Technologies - Volume 5: AI4Health*. 2018 Presented at: International Joint Conference on Biomedical Engineering Systems and Technologies; January 19, 2018; Maderia, Portugal p. 680-687 URL: <https://www.scitepress.org/Link.aspx?doi=10.5220/0006750606800687>
11. Musacchio N, Giancaterini A, Guaita G, Ozzello A, Pellegrini MA, Ponzani P, et al. Artificial intelligence and big data in diabetes care: a position statement of the Italian Association of Medical Diabetologists. *J Med Internet Res* 2020 Jun 22;22(6):e16922 [FREE Full text] [doi: [10.2196/16922](https://doi.org/10.2196/16922)] [Medline: [32568088](https://pubmed.ncbi.nlm.nih.gov/32568088/)]
12. Bauer GR, Lizotte DJ. Artificial intelligence, intersectionality, and the future of public health. *Am J Public Health* 2021 Jan;111(1):98-100. [doi: [10.2105/AJPH.2020.306006](https://doi.org/10.2105/AJPH.2020.306006)] [Medline: [33326280](https://pubmed.ncbi.nlm.nih.gov/33326280/)]
13. Timmins KA, Green MA, Radley D, Morris MA, Pearce J. How has big data contributed to obesity research? a review of the literature. *Int J Obes (Lond)* 2018 Dec 18;42(12):1951-1962 [FREE Full text] [doi: [10.1038/s41366-018-0153-7](https://doi.org/10.1038/s41366-018-0153-7)] [Medline: [30022056](https://pubmed.ncbi.nlm.nih.gov/30022056/)]
14. Lewis SP, Arbuthnott AE. Searching for thinspiration: the nature of internet searches for pro-eating disorder websites. *Cyberpsychol Behav Soc Netw* 2012 Apr;15(4):200-204. [doi: [10.1089/cyber.2011.0453](https://doi.org/10.1089/cyber.2011.0453)] [Medline: [22335543](https://pubmed.ncbi.nlm.nih.gov/22335543/)]
15. Oksanen A, Garcia D, Sirola A, Näsi M, Kaakinen M, Keipi T, et al. Pro-anorexia and anti-pro-anorexia videos on YouTube: sentiment analysis of user responses. *J Med Internet Res* 2015 Nov 12;17(11):e256 [FREE Full text] [doi: [10.2196/jmir.5007](https://doi.org/10.2196/jmir.5007)] [Medline: [26563678](https://pubmed.ncbi.nlm.nih.gov/26563678/)]
16. Fettach Y, Benhiba L. Pro-eating disorders and pro-recovery communities on Reddit: text and network comparative analyses. Presented at: 21st International Conference on Information Integration and Web-based Applications & Services; December 2, 2019; Munich, Germany p. 277-286. [doi: [10.1145/3366030.3366058](https://doi.org/10.1145/3366030.3366058)]
17. Benítez-Andrades J, Alija-Pérez J, García-Rodríguez I, Benavides C, Alaiz-Moretón H, Pastor-Vargas R, et al. BERT model-based approach for detecting categories of tweets in the field of eating disorders. 2021 Presented at: IEEE 34th International Symposium on Computer-Based Medical Systems; June 7, 2021; Aveiro, Portugal p. 586-590. [doi: [10.1109/cbms52027.2021.00105](https://doi.org/10.1109/cbms52027.2021.00105)]
18. Gamache R, Kharrazi H, Weiner J. Public and population health informatics: the bridging of big data to benefit communities. *Yearb Med Inform* 2018 Aug 29;27(1):199-206 [FREE Full text] [doi: [10.1055/s-0038-1667081](https://doi.org/10.1055/s-0038-1667081)] [Medline: [30157524](https://pubmed.ncbi.nlm.nih.gov/30157524/)]
19. Mei R. Health informatics and healthcare delivery: from the cost-effectiveness perspective. 2021 Apr 30 Presented at: 7th International Conference on Information Management; March 27, 2021; London, United Kingdom p. 62-65. [doi: [10.1109/icim52229.2021.9417045](https://doi.org/10.1109/icim52229.2021.9417045)]
20. Mheidly N, Fares J. Leveraging media and health communication strategies to overcome the COVID-19 infodemic. *J Public Health Policy* 2020 Dec 21;41(4):410-420 [FREE Full text] [doi: [10.1057/s41271-020-00247-w](https://doi.org/10.1057/s41271-020-00247-w)] [Medline: [32826935](https://pubmed.ncbi.nlm.nih.gov/32826935/)]
21. Zhang Z, Ahmed W. A comparison of information sharing behaviours across 379 health conditions on Twitter. *Int J Public Health* 2019 Apr;64(3):431-440 [FREE Full text] [doi: [10.1007/s00038-018-1192-5](https://doi.org/10.1007/s00038-018-1192-5)] [Medline: [30585297](https://pubmed.ncbi.nlm.nih.gov/30585297/)]
22. Kim A, Miano T, Chew R, Eggers M, Nonnemaker J. Classification of Twitter users who tweet about e-cigarettes. *JMIR Public Health Surveill* 2017 Sep 26;3(3):e63 [FREE Full text] [doi: [10.2196/publichealth.8060](https://doi.org/10.2196/publichealth.8060)] [Medline: [28951381](https://pubmed.ncbi.nlm.nih.gov/28951381/)]
23. Choi M, Kim S, Lee S, Kwon BC, Yi JS, Choo J, et al. Toward predicting social support needs in online health social networks. *J Med Internet Res* 2017 Aug 02;19(8):e272 [FREE Full text] [doi: [10.2196/jmir.7660](https://doi.org/10.2196/jmir.7660)] [Medline: [28768609](https://pubmed.ncbi.nlm.nih.gov/28768609/)]
24. Zhou S, Zhao Y, Bian J, Haynos AF, Zhang R. Exploring eating disorder topics on Twitter: machine learning approach. *JMIR Med Inform* 2020 Oct 30;8(10):e18273 [FREE Full text] [doi: [10.2196/18273](https://doi.org/10.2196/18273)] [Medline: [33124997](https://pubmed.ncbi.nlm.nih.gov/33124997/)]
25. Eysenbach G, Till JE. Ethical issues in qualitative research on internet communities. *BMJ* 2001 Nov 10;323(7321):1103-1105 [FREE Full text] [doi: [10.1136/bmj.323.7321.1103](https://doi.org/10.1136/bmj.323.7321.1103)] [Medline: [11701577](https://pubmed.ncbi.nlm.nih.gov/11701577/)]
26. Liu L, Woo BKP. Twitter as a mental health support system for students and professionals in the medical field. *JMIR Med Educ* 2021 Jan 19;7(1):e17598 [FREE Full text] [doi: [10.2196/17598](https://doi.org/10.2196/17598)] [Medline: [33464210](https://pubmed.ncbi.nlm.nih.gov/33464210/)]
27. Weder F, Krainer L, Karmasin M. *The Sustainability Communication Reader A Reflective Compendium*. Wiesbaden: Springer VS; 2021.
28. Géron A. Concepts, tools, and techniques to build intelligent systems. In: *Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow*, 2nd Edition. London, United Kingdom: O'Reilly Media; 2017.
29. Van Der Walt E, Eloff J. Using machine learning to detect fake identities: bots vs humans. *IEEE Access* 2018;6:6540-6549. [doi: [10.1109/access.2018.2796018](https://doi.org/10.1109/access.2018.2796018)]
30. Geirhos R, Meding K, Wichmann F. Beyond accuracy: quantifying trial-by-trial behaviour of CNNs and humans by measuring error consistency. 2020 Dec 06 Presented at: 34th Conference on Neural Information Processing Systems (NeurIPS 2020); December 6, 2020; Vancouver, Canada p. 200616736 URL: <http://arxiv.org/abs/2006.16736>

31. Youyou W, Kosinski M, Stillwell D. Computer-based personality judgments are more accurate than those made by humans. *Proc Natl Acad Sci U S A* 2015 Jan 27;112(4):1036-1040 [FREE Full text] [doi: [10.1073/pnas.1418680112](https://doi.org/10.1073/pnas.1418680112)] [Medline: [25583507](https://pubmed.ncbi.nlm.nih.gov/25583507/)]
32. A.Jabbar Alkubaisi GA, Kamaruddin SS, Husni H. Stock market classification model using sentiment analysis on Twitter based on hybrid naive bayes classifiers. *Comput Inf Sci* 2018 Jan 11;11(1):52. [doi: [10.5539/cis.v11n1p52](https://doi.org/10.5539/cis.v11n1p52)]
33. Prieto VM, Matos S, Álvarez M, Cacheda F, Oliveira JL. Twitter: a good place to detect health conditions. *PLoS One* 2014 Jan;9(1):e86191 [FREE Full text] [doi: [10.1371/journal.pone.0086191](https://doi.org/10.1371/journal.pone.0086191)] [Medline: [24489699](https://pubmed.ncbi.nlm.nih.gov/24489699/)]
34. Dixit A, Mani A, Bansal R. CoV2-Detect-Net: design of COVID-19 prediction model based on hybrid DE-PSO with SVM using chest X-ray images. *Inf Sci (N Y)* 2021 Sep;571:676-692 [FREE Full text] [doi: [10.1016/j.ins.2021.03.062](https://doi.org/10.1016/j.ins.2021.03.062)] [Medline: [33840820](https://pubmed.ncbi.nlm.nih.gov/33840820/)]
35. Singh V, Poonia RC, Kumar S, Dass P, Agarwal P, Bhatnagar V, et al. Prediction of COVID-19 corona virus pandemic based on time series data using support vector machine. *J Discrete Math Sci Crypto* 2020 Dec 14;23(8):1583-1597. [doi: [10.1080/09720529.2020.1784535](https://doi.org/10.1080/09720529.2020.1784535)]
36. Shofiya C, Abidi S. Sentiment analysis on COVID-19-related social distancing in Canada using Twitter data. *Int J Environ Res Public Health* 2021 Jun 03;18(11):5993 [FREE Full text] [doi: [10.3390/ijerph18115993](https://doi.org/10.3390/ijerph18115993)] [Medline: [34204907](https://pubmed.ncbi.nlm.nih.gov/34204907/)]
37. Yeom H, Hwang M, Hwang M, Jung H. study of machine-learning classifier and feature set selection for intent classification of Korean tweets about food safety. *J Inf Sci Theory Pract* 2014 Sep 30;2(3):29-39. [doi: [10.1633/jistap.2014.2.3.3](https://doi.org/10.1633/jistap.2014.2.3.3)]
38. Wang T, Brede M, Ianni A, Mentzakis E. Social interactions in online eating disorder communities: a network perspective. *PLoS One* 2018 Jul 30;13(7):e0200800 [FREE Full text] [doi: [10.1371/journal.pone.0200800](https://doi.org/10.1371/journal.pone.0200800)] [Medline: [30059512](https://pubmed.ncbi.nlm.nih.gov/30059512/)]
39. Wang SB. Machine learning to advance the prediction, prevention and treatment of eating disorders. *Eur Eat Disord Rev* 2021 Sep;29(5):683-691 [FREE Full text] [doi: [10.1002/erv.2850](https://doi.org/10.1002/erv.2850)] [Medline: [34231286](https://pubmed.ncbi.nlm.nih.gov/34231286/)]
40. Nadeem M. Identifying Depression on Twitter. arXiv. Preprint posted online on Jul 25, 2016. [FREE Full text]
41. Arseniev-Koehler A, Lee H, McCormick T, Moreno MA. #Proana: pro-eating disorder socialization on Twitter. *J Adolesc Health* 2016 Jun;58(6):659-664. [doi: [10.1016/j.jadohealth.2016.02.012](https://doi.org/10.1016/j.jadohealth.2016.02.012)] [Medline: [27080731](https://pubmed.ncbi.nlm.nih.gov/27080731/)]
42. Bert F, Gualano MR, Camussi E, Siliquini R. Risks and threats of social media websites: Twitter and the proana movement. *Cyberpsychol Behav Soc Netw* 2016 Apr;19(4):233-238. [doi: [10.1089/cyber.2015.0553](https://doi.org/10.1089/cyber.2015.0553)] [Medline: [26991868](https://pubmed.ncbi.nlm.nih.gov/26991868/)]
43. Branley DB, Covey J. Pro-ana versus pro-recovery: a content analytic comparison of social media users' communication about eating disorders on Twitter and Tumblr. *Front Psychol* 2017 Aug 11;8:1356 [FREE Full text] [doi: [10.3389/fpsyg.2017.01356](https://doi.org/10.3389/fpsyg.2017.01356)] [Medline: [28848472](https://pubmed.ncbi.nlm.nih.gov/28848472/)]
44. Kumar A, Singh JP, Dwivedi YK, Rana NP. A deep multi-modal neural network for informative Twitter content classification during emergencies. *Ann Oper Res* 2020 Jan 16. [doi: [10.1007/s10479-020-03514-x](https://doi.org/10.1007/s10479-020-03514-x)]
45. Madichetty S, Sridevi M. Classifying informative and non-informative tweets from the twitter by adapting image features during disaster. *Multimed Tools Appl* 2020 Aug 07;79(39-40):28901-28923. [doi: [10.1007/s11042-020-09343-1](https://doi.org/10.1007/s11042-020-09343-1)]
46. Congosto M, Basanta-Val P, Sanchez-Fernandez L. T-Hoarder: a framework to process Twitter data streams. *J Netw Comput Appl* 2017 Apr;83:28-39. [doi: [10.1016/j.jnca.2017.01.029](https://doi.org/10.1016/j.jnca.2017.01.029)]
47. simpletransformers. Github. URL: <https://github.com/ThilinaRajapakse/simpletransformers> [accessed 2021-10-26]
48. Liu Y, Ott M, Goyal N. RoBERTa: a robustly optimized BERT pretraining approach. arXiv. Preprint posted online on Jul 26, 2019. [FREE Full text]
49. Devlin J, Chang M, Lee K, Toutanova K. BERT: pre-training of deep bidirectional transformers for language understanding. arXiv. Preprint posted online on May 24, 2019. [FREE Full text]
50. Martin L, Muller B, Suárez P. CamemBERT: a tasty french language model. 2020 Presented at: 58th Annual Meeting of the Association for Computational Linguistics; July 1, 2020; Online p. 7203-7219. [doi: [10.18653/v1/2020.acl-main.645](https://doi.org/10.18653/v1/2020.acl-main.645)]
51. Sanh V, Debut L, Chaumond J, Wolf T. DistilBERT, a distilled version of BERTmaller, faster, cheaper and lighter. 2019 Presented at: Fifth Workshop on Energy Efficient Machine Learning and Cognitive Computing; December 13, 2019; Vancouver, BC p. 191001108 URL: <https://www.emc2-ai.org/assets/docs/neurips-19/emc2-neurips19-paper-33.pdf>
52. Le H, Vial L, Frej J. FlauBERT: unsupervised language model pre-training for french. arXiv. Preprint posted online on December 11, 2019. [FREE Full text]
53. Lan Z, Chen M, Goodman S, Gimpel K, Sharma P, Soricut R. ALBERT: a lite BERT for self-supervised learning of language representations. arXiv. Preprint posted online on Feb 9, 2020. [FREE Full text]
54. Delobelle P, Winters T, Berendt B. RobBERT: a Dutch RoBERTa-based language model. 2020 Nov 01 Presented at: Association for Computational Linguistics: EMNLP 2020; November 2020; Online p. 3255-3265 URL: <https://aclanthology.org/2020.findings-emnlp.292/> [doi: [10.18653/v1/2020.findings-emnlp.292](https://doi.org/10.18653/v1/2020.findings-emnlp.292)]
55. Saon G, Tüske Z, Bolanos D, Kingsbury B. Advancing RNN transducer technology for speech recognition. 2021 Presented at: 2021 IEEE International Conference on Acoustics, Speech and Signal Processing; June 6-11, 2021; Toronto, ON, Canada. [doi: [10.1145/3437963.3441667](https://doi.org/10.1145/3437963.3441667)]
56. Enhancing BERT for lexical normalization. HAL-Inria. URL: <https://hal.inria.fr/hal-02294316/> [accessed 2022-02-18]
57. Shahid F, Zameer A, Muneeb M. Predictions for COVID-19 with deep learning models of LSTM, GRU and Bi-LSTM. *Chaos Solitons Fractals* 2020 Nov;140:110212 [FREE Full text] [doi: [10.1016/j.chaos.2020.110212](https://doi.org/10.1016/j.chaos.2020.110212)] [Medline: [32839642](https://pubmed.ncbi.nlm.nih.gov/32839642/)]
58. Breiman L. Random forests. *Machine Learning* 2001;45:5-32 [FREE Full text] [doi: [10.1023/A:1010933404324](https://doi.org/10.1023/A:1010933404324)]

59. Chen X, Wu Y, Wang Z, Liu S, Li J. Developing real-time streaming transformer transducer for speech recognition on large-scale dataset. 2021 Presented at: 2021 IEEE International Conference on Acoustics, Speech and Signal Processing; June 6-11, 2021; Toronto, Canada.
60. Hamdi Y, Boubaker H, Alimi AM. Online Arabic handwriting recognition using graphemes segmentation and deep learning recurrent neural networks. In: Hassanien AE, Darwish A, Abd El-Kader SM, Alboaneen DA, editors. Enabling Machine Learning Applications in Data Science. Algorithms for Intelligent Systems. Singapore: Springer; 2021.
61. Hochreiter S, Schmidhuber J. Long short-term memory. *Neural Comput* 1997 Nov 15;9(8):1735-1780. [doi: [10.1162/neco.1997.9.8.1735](https://doi.org/10.1162/neco.1997.9.8.1735)] [Medline: [9377276](https://pubmed.ncbi.nlm.nih.gov/9377276/)]
62. Yates A, Nogueira R, Lin J. Pretrained transformers for text ranking: BERT and beyond. 2021 Presented at: 14th ACM International Conference on Web Search and Data Mining; March 8-11; Israel (online). [doi: [10.1145/3437963.3441667](https://doi.org/10.1145/3437963.3441667)]
63. Benítez-Andrades J. jabenitez88/NLP-EatingDisordersBERT: Categorizing tweets about eating disorders using text mining: BERT models and machine learning techniques. Zenodo. 2021 Jul 31. URL: [https://zenodo.org/record/5148631#.YhU1tJPMI\\_U](https://zenodo.org/record/5148631#.YhU1tJPMI_U) [accessed 2022-02-22]
64. Benítez-Andrades JA. Eating disorders tweets. Kaggle. URL: <https://www.kaggle.com/jabenitez88/eating-disorders-tweets/activity> [accessed 2021-12-26]
65. Roitero K, Bozzato C, Mea V, Mizzaro S, Serra G. Twitter goes to the doctor: detecting medical tweets using machine learning and BERT. CEUR Workshop. 2020 Jan 01. URL: <http://ceur-ws.org/Vol-2619/short1.pdf> [accessed 2022-02-22]

*Edited by C Lovis; submitted 03.12.21; peer-reviewed by L Prieto Santamaría, J Rafael, EJ González González, Anonymous, J Wang, SC Lu; comments to author 24.12.21; revised version received 07.01.22; accepted 01.02.22; published 24.02.22.*

*Please cite as:*

*Benítez-Andrades JA, Alija-Pérez JM, Vidal ME, Pastor-Vargas R, García-Ordás MT*

*Traditional Machine Learning Models and Bidirectional Encoder Representations From Transformer (BERT)-Based Automatic Classification of Tweets About Eating Disorders: Algorithm Development and Validation Study*

*JMIR Med Inform* 2022;10(2):e34492

URL: <https://medinform.jmir.org/2022/2/e34492>

doi:[10.2196/34492](https://doi.org/10.2196/34492)

PMID:[35200156](https://pubmed.ncbi.nlm.nih.gov/35200156/)

©José Alberto Benítez-Andrades, José-Manuel Alija-Pérez, María-Esther Vidal, Rafael Pastor-Vargas, María Teresa García-Ordás. Originally published in JMIR Medical Informatics (<https://medinform.jmir.org>), 24.02.2022. This is an open-access article distributed under the terms of the Creative Commons Attribution License (<https://creativecommons.org/licenses/by/4.0/>), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Medical Informatics, is properly cited. The complete bibliographic information, a link to the original publication on <https://medinform.jmir.org/>, as well as this copyright and license information must be included.

Original Paper

# Identification of Prediabetes Discussions in Unstructured Clinical Documentation: Validation of a Natural Language Processing Algorithm

Jessica L Schwartz<sup>1,2\*</sup>, MD, MHS; Eva Tseng<sup>1,3\*</sup>, MD, MPH; Nisa M Maruthur<sup>1,3,4\*</sup>, MD, MHS; Masoud Rouhizadeh<sup>5,6\*</sup>, MS, PhD

<sup>1</sup>Division of General Internal Medicine, Johns Hopkins School of Medicine, Baltimore, MD, United States

<sup>2</sup>Division of Hospital Medicine, Johns Hopkins Hospital, Baltimore, MD, United States

<sup>3</sup>Welch Center for Prevention, Epidemiology, & Clinical Research, Johns Hopkins University, Baltimore, MD, United States

<sup>4</sup>Department of Epidemiology, Johns Hopkins University Bloomberg School of Public Health, Baltimore, MD, United States

<sup>5</sup>Department of Pharmaceutical Outcomes and Policy, University of Florida College of Pharmacy, Gainesville, FL, United States

<sup>6</sup>Division of Biomedical Informatics and Data Science, Johns Hopkins University School of Medicine, Baltimore, MD, United States

\* all authors contributed equally

**Corresponding Author:**

Jessica L Schwartz, MD, MHS

Division of General Internal Medicine

Johns Hopkins School of Medicine

2024 E Monument St.

Ste 2-604D

Baltimore, MD, 21205

United States

Phone: 1 973 722 8552

Fax: 1 410 955 0476

Email: [jschwa64@jhmi.edu](mailto:jschwa64@jhmi.edu)

## Abstract

**Background:** Prediabetes affects 1 in 3 US adults. Most are not receiving evidence-based interventions, so understanding how providers discuss prediabetes with patients will inform how to improve their care.

**Objective:** This study aimed to develop a natural language processing (NLP) algorithm using machine learning techniques to identify discussions of prediabetes in narrative documentation.

**Methods:** We developed and applied a keyword search strategy to identify discussions of prediabetes in clinical documentation for patients with prediabetes. We manually reviewed matching notes to determine which represented actual prediabetes discussions. We applied 7 machine learning models against our manual annotation.

**Results:** Machine learning classifiers were able to achieve classification results that were close to human performance with up to 98% precision and recall to identify prediabetes discussions in clinical documentation.

**Conclusions:** We demonstrated that prediabetes discussions can be accurately identified using an NLP algorithm. This approach can be used to understand and identify prediabetes management practices in primary care, thereby informing interventions to improve guideline-concordant care.

(*JMIR Med Inform* 2022;10(2):e29803) doi:[10.2196/29803](https://doi.org/10.2196/29803)

**KEYWORDS**

prediabetes; prediabetes discussions; prediabetes management; chronic disease management; physician-patient communication; natural language processing; machine learning

## Introduction

Prediabetes affects 88 million US adults [1,2], and evidence-based interventions focusing on lifestyle modification can prevent type 2 diabetes [3-12]. In particular, the Diabetes Prevention Program is an effective lifestyle intervention that decreases diabetes incidence, with the most recent data showing a 27% risk reduction compared with the placebo arm over 15 years of follow up [5]. Unfortunately, up to 89% of patients do not know they have prediabetes [13,14], and many patients are unaware of interventions to decrease their risk of diabetes—relying on their primary care providers (PCPs) to initiate discussions about diabetes prevention, including the importance of lifestyle changes [8,9]. However, survey data demonstrate that many providers feel that they lack the resources to effectively implement evidence-based prediabetes treatment [8,9]. Focused primary care interventions to support decision-making and education may be able to improve diagnosis of prediabetes and delivery of guideline-concordant care.

Rigorous quality improvement interventions require evaluation using measurement before and after implementation of a project to determine whether there is a demonstrable change in target outcomes. Unfortunately, it is difficult to identify changes and improvement in prediabetes management through structured data alone. Relying on diagnosis codes is insufficient; one study showed that only 13% of patients with prediabetes had an International Classification of Diseases (ICD)-9 diagnosis of prediabetes or hyperglycemia [14]. Although labs, orders, and referrals provide some insight, this information lacks detail about management, particularly lifestyle counseling, which is better captured in narrative documentation. This content is not easily queried and requires innovative research methods to accurately reflect delivery of prediabetes care.

Prior studies have shown that natural language processing (NLP) can be used to diagnose chronic conditions, like diabetes, but few focus on disease management [15]. Similarly, NLP studies in prediabetes have primarily focused on disease detection, screening, and predictive modeling, with no studies applying machine learning (ML) techniques to determine how prediabetes is managed [16-27]. Our goal was to develop a method to identify when providers discuss prediabetes management and treatment, which could later be used to determine if care delivered meets evidence-based guidelines and compare outcomes before and after an intervention. Therefore, we developed and validated NLP pipelines to identify primary care discussions about prediabetes in clinical documentation.

## Methods

### Population and Ethics Approval

We identified patients with prediabetes who had an internal medicine primary care visit within an academic center with multiple ambulatory locations in Maryland and Washington, DC. Eligible patients were adults ( $\geq 18$  years old) covered by 1 of 3 major insurers who completed an in-person visit and had a hemoglobin A<sub>1c</sub> (HbA<sub>1c</sub>) level between 5.7% and 6.4%

between July 1, 2016 and December 31, 2018. Patients with diabetes (any type) based on billing codes or documentation in the problem list or past medical history were excluded. Data cleaning and analyses were performed using Stata 15. This study was approved by the Johns Hopkins Institutional Review Board (IRB00196984).

### Keyword Search Refinement (Phase 1)

Based on clinical experience, we developed a list of keywords used to describe “prediabetes” (Table S1 in [Multimedia Appendix 1](#)). We identified visit notes containing these keywords using Python string matching and dictionary look-up, accounting for variations like spelling errors and morphological differences. We extracted a  $\pm 25$ -word concordance window (“note snippet”) for each match to provide textual context. Multiple snippets could come from the same note if multiple matching keywords were present.

We selected 2 ambulatory clinics from our overall population. Of 315 patients meeting inclusion criteria, 40.6% (128/315) had at least one matching keyword during the study period. These patients had a total of 637 keyword matches across 324 encounters with 25 providers. We conducted manual annotation to determine which of the 637 note snippets represented true clinical discussions of prediabetes (yes or no). Outpatient provider documentation typically includes chief complaint, history of present illness, medical and family history, objective data including physical exam, and an assessment and plan. We considered use of a section identification pipeline to exclude specific sections of the notes (eg, past medical history) in which keywords would not represent prediabetes discussions. However, section identification pipelines are less generalizable, and the providers in our sample did not use standardized templates, making section boundaries difficult to define [28]. Instead, note snippets were designated “no” during manual review if the keyword was only present in past medical history, a list of diagnoses outside of the assessment and plan, family history, or the description of a lab result.

We double-reviewed a random sample of 200 note snippets. Interrater reliability (IRR) was 95%. Discrepancies between annotators were resolved via consensus to refine the definition of “prediabetes discussion.” We then manually reviewed patient records for 35.3% (66/187) of charts without a keyword match to identify false negatives. We reviewed all notes written by the patient’s PCP within the inclusion timeframe, and 9% (6/66) of patients had prediabetes discussions that were not captured. We added 3 keywords (“dysglycemia,” “hyperglycemia,” and “pre diabetes”) to the lexicon (Table S1 in [Multimedia Appendix 1](#)).

### Training Set (Phase 2)

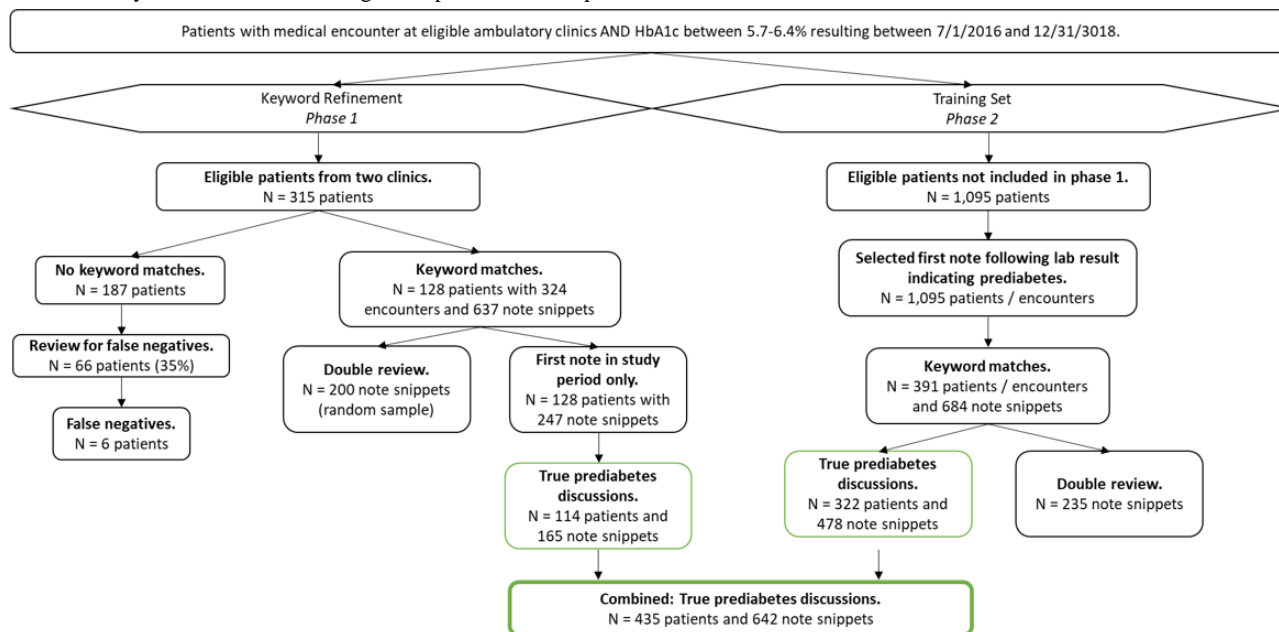
We developed a training set to test our prediabetes lexicon against patients from clinics not included in phase 1 ([Figure 1](#)). We included a single note per patient (n=1095), choosing the first encounter after the HbA<sub>1c</sub> result that met inclusion criteria. We applied the finalized keyword search, which resulted in 684 matches for 381 patients seen by 73 providers. We abstracted the 684 note snippets and annotated the notes using a similar process as above. We double-reviewed 34% of the note snippets



with an IRR of 97% for manual annotation, resolving to 100% agreement upon review. We combined these results with note snippets from phase 1. To avoid overselection of a single patient

or provider, we included note snippets from 1 encounter per patient for a total of 930 note snippets written by 96 unique providers.

**Figure 1.** Diagram depicting selection and review during keyword search refinement (Phase 1) and training set development (Phase 2). Eligible patients were adults ( $\geq 18$  years old) covered by 1 of 3 major insurers who completed an in-person visit at a Johns Hopkins clinic and had an HbA<sub>1c</sub> level between 5.7% and 6.4% (39-46 mmol/mol) between July 1, 2016 and December 31, 2018. Note, double review indicates that 2 providers reviewed the keyword matches to identify whether the surrounding text represented a true prediabetes discussion.



## Rule-Based System

Rule-based systems are frequently used for clinical concept extraction and text classification systems because of their ease of implementation and minimal computational requirements. To establish a strong baseline, we tested the feasibility of identifying prediabetes discussions with a rule-based classification scheme. Using the spaCy EntityRuler module [29], we created 42 expert-developed patterns that, if present, would represent prediabetes discussions. The spaCy EntityRuler module facilitates various pattern, keyword, and regular expression searching and matching and allows us to account for morphological variations (eg, singular vs plural forms, conjunctions), as well as substitutions of different prepositions (eg, about vs for) and synonyms (eg, prediabetes, impaired fasting glucose). Table S2 in [Multimedia Appendix 1](#) provides our expert-developed patterns for this rule-based system. We randomly sampled 90% of the note snippets to develop and revise the rule-based system and evaluated the system on the remaining 10%.

## Machine Learning

### Feature Selection

Note snippets from the training set were stemmed using the Porter stemmer, and common stop words were removed using the Natural Language Toolkit (NLTK) stop word list [30]. We used the Python scikit-learn library [31] to extract word ngram sequences (1-5 grams), weighted by term frequency-inverse document frequency (TF-IDF) [32]. We applied logistic regression with L1 regularization [33] to reduce the dimensionality of the feature vectors.

## Computational Environment

Deep learning and ML experiments were conducted on the Johns Hopkins University (JHU) Precision Medicine Analytics Platform (PMAP), a high-performance, cloud-based, big-data platform to accelerate biomedical discovery and translate discovered knowledge to improve patient-centered care. PMAP pulls data from the Johns Hopkins Medicine electronic health record (EHR) to support processing by ML and NLP technologies. Statistical analysis and manual annotation were done in the JHU Secure Analytic Framework Environment, a virtual desktop that provides JHU investigators with a secure platform for analyzing and sharing sensitive data (including protected health information) with colleagues.

## Classification

We used the labeled note snippets to train multiple ML classifiers to replicate human annotation for prediabetes discussions. We applied 6 binary classification models: logistic regression [34], linear support vector machines (SVM) [35], stochastic gradient descent (SGD) [36], decision tree [37], random forest [38,39], and Gaussian naïve Bayes (NB) [40]. To reduce overfitting, each model was evaluated using 10-fold cross-validation by training, randomly, on 90% of the data and holding out 10% for testing. All modeling was performed in scikit-learn [31].

We also applied convolutional neural networks (CNNs) for sentence categorization [41], a well-established deep learning method in NLP for text classification [42] using Python spaCy 2.1 implementation [29]. We started with the tokenization of each note snippet and creating an embedding vector of each token using scispaCy large models (~785,000 vocabulary and

600,000 word vectors), pretrained on biomedical and clinical text [43]. Next, to represent the tokens in context, these vectors were encoded into a sentence matrix by computing the vector for each token using a forward pass and a backward pass. After that, a self-attention mechanism was applied to reduce the dimensionality of the sentence matrix representation into a single context vector. Finally, these vectors were average-pooled and used as features in a simple feed-forward network for predicting true discussions of prediabetes. For the CNN model, we used the spaCy 2.2 default network architecture and parameters [44].

For each classification method, we reported on agreement, sensitivity and recall, specificity, positive predictive value and precision, and F measure using manual annotation as the gold standard. To test statistical significance between classification methods, we used MLxtend Python library to perform a 5x2 cross-validation paired *t* test [45]. A *P* value <.05 indicated that

we could reject the null hypothesis that both models performed equally to classify prediabetes discussions.

## Results

We identified 1410 patients with prediabetes; 518 (36.74%) had at least one keyword match. Among these patients, 435 (84.0%) had a true discussion about prediabetes in the manually reviewed documents (Figure 1).

The rule-based system was inadequate for replicating human performance, with 72.5% recall and 42.6% specificity (Table 1). ML and CNN classification, however, were close to human performance across all models (Table 1). When comparing conventional classifiers with logistic regression (which had the highest agreement), only linear SVM and NB had similar performance ( $P=.11$  and  $P=.15$ , respectively). CNN outperformed all conventional ML classifiers (logistic regression:  $P=.04$ ; SVM:  $P=.02$ ; SGD:  $P=.002$ ; random forest:  $P=.002$ ; decision tree:  $P=.001$ ; NB:  $P=.03$ ).

**Table 1.** Performance of machine learning methods to approximate manual annotation in identifying prediabetes discussions from primary care note snippets (n=930).

Method	Instances classifier agreed with manual annotation, n (%)	Sensitivity/recall	Specificity	PPV <sup>a</sup> /precision	F measure
<b>Rule-based system</b>					
Expert-developed patterns	588 (63.2)	0.725	0.426	0.737	0.731
<b>Machine learning</b>					
Logistic regression	885 (95.2)	0.966	0.921	0.965	0.965
Linear support vector machines	878 (94.4)	0.962	0.903	0.957	0.960
Stochastic gradient descent	858 (92.3)	0.926	0.915	0.96	0.943
Random forest	863 (92.8)	0.961	0.854	0.937	0.948
Decision tree	832 (89.5)	0.923	0.83	0.925	0.924
Gaussian naïve Bayes	883 (95.0)	0.966	0.912	0.96	0.963
Convolutional neural networks	910 (97.9)	0.984	0.966	0.984	0.984

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

Manual annotation revealed a variety of linguistic patterns that did and did not represent clinical discussions of prediabetes (Table 2). Most commonly, true discussions were found in the assessment and plan, and those that did not were auto populated

from structured fields. ML did result in 5% misclassification based on logistic regression, the best performing conventional classifier; a pattern was not apparent on review of these misclassified note snippets.

**Table 2.** Example text from clinical documentation containing keywords matching the “prediabetes” extraction lexicon, stratified by whether the text represents documentation of a prediabetes discussion.

Location in note	Representative text from note snippets <sup>a</sup>
<b>Text containing keyword matches representing prediabetes discussions.</b>	
Chief complaint	<ul style="list-style-type: none"> <li>Chief complaint: Patient is a 42 y.o. female here with questions about prediabetes.</li> <li>Patient presents to the visit for an annual physical and reevaluation of HTN<sup>b</sup> and impaired fasting glucose.</li> </ul>
History of Present Illness	<ul style="list-style-type: none"> <li>Has a treadmill but not using regularly. Recent a1c was 6.2 consistent with pre-diabetes.</li> </ul>
Visit Problem List	<ul style="list-style-type: none"> <li>Problem List Items Addressed This Visit Asthma Borderline diabetes Essential hypertension</li> <li>Assessment Order Plan 1. Hyperlipidemia ... 7. Impaired fasting glucose 8. Health care maintenance</li> </ul>
Assessment & Plan	<ul style="list-style-type: none"> <li>Hyperglycemia Lifestyle modification including diet and exercise discussed. 6. Elevated blood pressure.</li> <li>Pre-diabetes Assessment: recent A1C in good range. Plan: exercise and healthy food changes.</li> </ul>
<b>Text containing keyword matches not representing prediabetes discussions.</b>	
One-liner	<ul style="list-style-type: none"> <li>Patient with history of HTN, HLD<sup>c</sup>, prediabetes, scleroderma here for routine health assessment.</li> </ul>
Past Medical History	<ul style="list-style-type: none"> <li>Past Medical History: Diagnosis Date Asthma 5/14/2008 ... Prediabetes 2/6/2012 Osteoporosis 5/14/2008</li> </ul>
Problem List	<ul style="list-style-type: none"> <li>... Hyperlipidemia E78.5 Impaired fasting glucose R73.01 Overweight E66.3 ...</li> </ul>
Diagnosis list	<ul style="list-style-type: none"> <li>Diagnoses of Essential hypertension, Osteoporosis, ..., Prediabetes, Asthma, ...</li> </ul>
Family history	<ul style="list-style-type: none"> <li>Family History Problem Relation Age of Onset Diabetes Father Prediabetes Paternal Grandfather...</li> </ul>
Pertinent positive	<ul style="list-style-type: none"> <li>Diagnosis remains unclear. He has prediabetes. Reports 2-3 months of intermittent palpitations.</li> </ul>
Pertinent negative	<ul style="list-style-type: none"> <li>Likely has peripheral neuropathy. Negative RPR<sup>d</sup>, HIV, pre-diabetes.</li> </ul>
Follow up reasons	<ul style="list-style-type: none"> <li>Follow up in 1 month for flu shot and prediabetes discussion.</li> </ul>
Results <sup>e</sup>	<ul style="list-style-type: none"> <li>For someone without known diabetes, a hemoglobin A<sub>1c</sub> value between 5.7 % and 6.4 % is consistent with prediabetes and should be confirmed.</li> </ul>
General guidelines <sup>e</sup>	<ul style="list-style-type: none"> <li>Type 2 diabetes or prediabetes All men beginning at age 45 and men without symptoms at any age who are overweight or obese and have 1 or more other risk factors.</li> </ul>

<sup>a</sup>Text was modified for length and content to serve as general examples while protecting patient anonymity.

<sup>b</sup>HTN: hypertension.

<sup>c</sup>HLD: hyperlipidemia.

<sup>d</sup>RPR: rapid plasma reagin.

<sup>e</sup>Populated in notes from clinical decision support tools.

## Discussion

### Principal Findings

We utilized NLP and ML techniques to identify prediabetes discussions from unstructured narrative documentation with up to 98% precision and recall. To date, NLP techniques have been used in prediabetes for screening, diagnosis, risk stratification, predictive modeling, and intervention design [16-27,46-50]. To our knowledge, this is the first NLP tool to identify prediabetes discussions. NLP methods have been applied in health care in many ways including in EHR free-text clinical notes to classify disease phenotype, with most studies using simple methods like shallow classifiers or combined with rule-based methods [15,51]. Compared with these studies, our NLP methods are not novel, but our application to disease management distinguishes our

study from those that primarily focus on condition identification for chronic diseases [15].

In our study, a simple rule-based system was inadequate to identify prediabetes discussions due to poor specificity. In contrast, all ML methods performed well, with 89% to 98% accuracy. This result demonstrates that prediabetes discussions, despite a variety of documentation styles, can be identified using NLP pipelines. Logistic regression, an efficient conventional classifier with minimal technical dependencies, was statistically outperformed by CNN, a deep learning technique. However, both identified >95% of prediabetes discussions, suggesting that either method could be applied depending on system needs.

Our NLP tool has multiple applications. The simplicity of logistic regression allows for deployment in operational settings,

particularly clinical decision support. The tool can also simplify the analytic process before and after a clinical intervention intended to change provider practices. For example, it can isolate discussions about prediabetes, a task that otherwise requires time-consuming manual review. The context of these discussions could then be reviewed to understand the impact of an intervention. This process would strengthen the evaluation of quality improvement programs for prediabetes to promote guideline-concordant care, which includes lifestyle counseling [3-7]. These methods should be replicable to identify conversations about behavioral interventions for other conditions, such as obesity, polysubstance abuse, or tobacco use, that rely heavily on counseling in addition to medication management and referrals.

### Strengths

Our study has several strengths. The keyword refinement stage was rigorous. We validated the initial keyword list against a random sample from 2 ambulatory clinics, ensuring we reviewed a variety of documentation styles. Manual annotation was performed by 2 experts to standardize our definition of “prediabetes discussion,” leading to improvement in IRR scores during training set development. We also identified false negatives and revised our initial keyword list accordingly to ensure capture of prediabetes discussions. Finally, we applied the search criteria developed during keyword refinement to a new set of notes from unique clinics to reduce overfitting. There was a total of 96 different providers included in the 930 unique note snippets, which allowed the model to learn the vocabulary and writing styles of many different clinicians.

### Limitations

Limitations of our study include collection of data from a single health system. However, the clinics included represent urban and suburban sites serving patients of different socioeconomic

levels and disease burden, improving generalizability. Providers at other institutions may use different medical terminology, not considered in this study, to describe “prediabetes.” This could limit generalizability outside of the home-trained institution. However, we took several steps to reduce institutional bias, including rigorous keyword refinement and application of the final lexical search to multiple clinics that do not share standardized templates to include many linguistic styles and patterns. We limited our note selection to the first encounter following the abnormal HbA<sub>1c</sub> result; although this could miss some dialogue about prediabetes, logically these discussions are most likely to occur close to the time of the abnormal result, and this decreased bias in our models. Finally, the note selection process, requiring at least one prediabetes keyword to enter our data set, limited our ability to calculate true recall. We minimized this issue by performing manual review on a subset of the charts that did not enter our data set, to ensure we did not have selection bias in our keyword search. Future studies may consider applying our NLP pipeline against a random sample of notes without requiring keyword selection to perform additional validations. Additionally, our study provides a baseline framework for identifying discussions of prediabetes. Next steps could apply NLP pipelines to identify when discussions about prediabetes meet the threshold for delivery of guideline-concordant care.

### Conclusion

Our NLP pipeline successfully identified prediabetes discussions in unstructured notes with precision approximating human annotation. This approach can be used to evaluate prediabetes counseling during patient visits and describe prediabetes management in primary care. Gathering these data is a critical step to inform interventions to improve the delivery of evidence-based prediabetes care to reduce the incidence of type 2 diabetes.

---

### Acknowledgments

This work was supported by the Johns Hopkins Institute for Clinical and Translational Research Core Coins Award 2018. ET was supported by the National Institute of Diabetes and Digestive and Kidney Diseases [K23DK118205]. JLS was supported by the National Heart, Lung, and Blood Institute [5T32HL007180, PI: Hill-Briggs].

---

### Conflicts of Interest

NMM is the co-inventor of a virtual diabetes prevention program. Under a license agreement between Johns Hopkins HealthCare Solutions and the Johns Hopkins University, NMM and the University are entitled to royalty distributions related to this technology. This arrangement has been reviewed and approved by the Johns Hopkins University in accordance with its conflict of interest policies. This technology is not described in this study. JLS is a co-investigator on a research project funded by NovoNordisk Inc. The primary aim of the project is to create and pilot a clinical decision support tool to assist clinicians when talking to their patients about weight and obesity treatment. This project is not addressed or referenced in this publication.

---

### Multimedia Appendix 1

Supplementary methods and tables.

[[DOCX File, 30 KB](#) - [medinform\\_v10i2e29803\\_app1.docx](#) ]

---

### References

1. National Diabetes Statistics Report, 2020: Estimates of Diabetes and Its Burden in the United States. Centers for Disease Control and Prevention. URL: <https://www.cdc.gov/diabetes/pdfs/data/statistics/national-diabetes-statistics-report.pdf> [accessed 2022-01-31]
2. National Diabetes Statistics Report: Estimates of Diabetes and Its Burden in the United States. Centers for Disease Control and Prevention. 2014. URL: <https://www.cdc.gov/diabetes/data/statistics-report/index.html> [accessed 2022-01-31]
3. Diabetes Prevention Program Research Group, Knowler WC, Fowler SE, Hamman RF, Christophi CA, Hoffman HJ, et al. 10-year follow-up of diabetes incidence and weight loss in the Diabetes Prevention Program Outcomes Study. *Lancet* 2009 Dec 14;374(9702):1677-1686 [FREE Full text] [doi: [10.1016/S0140-6736\(09\)61457-4](https://doi.org/10.1016/S0140-6736(09)61457-4)] [Medline: [19878986](https://pubmed.ncbi.nlm.nih.gov/19878986/)]
4. Knowler WC, Barrett-Connor E, Fowler SE, Hamman RF, Lachin JM, Walker EA, Diabetes Prevention Program Research Group. Reduction in the incidence of type 2 diabetes with lifestyle intervention or metformin. *N Engl J Med* 2002 Mar 07;346(6):393-403 [FREE Full text] [doi: [10.1056/NEJMoa012512](https://doi.org/10.1056/NEJMoa012512)] [Medline: [11832527](https://pubmed.ncbi.nlm.nih.gov/11832527/)]
5. 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 Dec;3(11):866-875 [FREE Full text] [doi: [10.1016/S2213-8587\(15\)00291-0](https://doi.org/10.1016/S2213-8587(15)00291-0)] [Medline: [26377054](https://pubmed.ncbi.nlm.nih.gov/26377054/)]
6. American Diabetes Association. 3. Prevention or Delay of Type 2 Diabetes. *Diabetes Care* 2020 Jan;43(Suppl 1):S32-S36. [doi: [10.2337/dc20-S003](https://doi.org/10.2337/dc20-S003)] [Medline: [31862746](https://pubmed.ncbi.nlm.nih.gov/31862746/)]
7. National Diabetes Prevention Program. Centers for Disease Control and Prevention. 2017. URL: <http://www.cdc.gov/diabetes/prevention/index.htm> [accessed 2022-01-31]
8. Tseng E, Greer RC, O'Rourke P, Yeh H, McGuire MM, Clark JM, et al. Survey of primary care providers' knowledge of screening for, diagnosing and managing prediabetes. *J Gen Intern Med* 2017 Dec;32(11):1172-1178 [FREE Full text] [doi: [10.1007/s11606-017-4103-1](https://doi.org/10.1007/s11606-017-4103-1)] [Medline: [28730532](https://pubmed.ncbi.nlm.nih.gov/28730532/)]
9. Tseng E, Greer RC, O'Rourke P, Yeh H, McGuire MM, Albright AL, et al. National survey of primary care physicians' knowledge, practices, and perceptions of prediabetes. *J Gen Intern Med* 2019 Nov;34(11):2475-2481 [FREE Full text] [doi: [10.1007/s11606-019-05245-7](https://doi.org/10.1007/s11606-019-05245-7)] [Medline: [31502095](https://pubmed.ncbi.nlm.nih.gov/31502095/)]
10. Rhee MK, Herrick K, Ziemer DC, Vaccarino V, Weintraub WS, Narayan KMV, et al. Many Americans have pre-diabetes and should be considered for metformin therapy. *Diabetes Care* 2010 Jan;33(1):49-54 [FREE Full text] [doi: [10.2337/dc09-0341](https://doi.org/10.2337/dc09-0341)] [Medline: [19808929](https://pubmed.ncbi.nlm.nih.gov/19808929/)]
11. Karve A, Hayward RA. Prevalence, diagnosis, and treatment of impaired fasting glucose and impaired glucose tolerance in nondiabetic U.S. adults. *Diabetes Care* 2010 Dec;33(11):2355-2359 [FREE Full text] [doi: [10.2337/dc09-1957](https://doi.org/10.2337/dc09-1957)] [Medline: [20724649](https://pubmed.ncbi.nlm.nih.gov/20724649/)]
12. Moin T, Li J, Duru OK, Ettner S, Turk N, Keckhafer A, et al. Metformin prescription for insured adults with prediabetes from 2010 to 2012: a retrospective cohort study. *Ann Intern Med* 2015 May 21;162(8):542-548 [FREE Full text] [doi: [10.7326/M14-1773](https://doi.org/10.7326/M14-1773)] [Medline: [25894024](https://pubmed.ncbi.nlm.nih.gov/25894024/)]
13. Centers for Disease Control and Prevention (CDC). Awareness of prediabetes--United States, 2005-2010. *MMWR Morb Mortal Wkly Rep* 2013 Mar 22;62(11):209-212 [FREE Full text] [Medline: [23515058](https://pubmed.ncbi.nlm.nih.gov/23515058/)]
14. Schmittiel JA, Adams SR, Segal J, Griffin MR, Roumie CL, Ohnsorg K, et al. Novel use and utility of integrated electronic health records to assess rates of prediabetes recognition and treatment: brief report from an integrated electronic health records pilot study. *Diabetes Care* 2014 Mar;37(2):565-568 [FREE Full text] [doi: [10.2337/dc13-1223](https://doi.org/10.2337/dc13-1223)] [Medline: [24271190](https://pubmed.ncbi.nlm.nih.gov/24271190/)]
15. Sheikhalishahi S, Miotto R, Dudley JT, Lavelli A, Rinaldi F, Osmani V. Natural language processing of clinical notes on chronic diseases: systematic review. *JMIR Med Inform* 2019 May 27;7(2):e12239 [FREE Full text] [doi: [10.2196/12239](https://doi.org/10.2196/12239)] [Medline: [31066697](https://pubmed.ncbi.nlm.nih.gov/31066697/)]
16. De Silva K, Jönsson D, Demmer RT. A combined strategy of feature selection and machine learning to identify predictors of prediabetes. *J Am Med Inform Assoc* 2020 Mar 01;27(3):396-406 [FREE Full text] [doi: [10.1093/jamia/ocz204](https://doi.org/10.1093/jamia/ocz204)] [Medline: [31889178](https://pubmed.ncbi.nlm.nih.gov/31889178/)]
17. Chung JW, Kim WJ, Choi SB, Park JS, Kim DW. Screening for pre-diabetes using support vector machine model. *Annu Int Conf IEEE Eng Med Biol Soc* 2014;2014:2472-2475. [doi: [10.1109/EMBC.2014.6944123](https://doi.org/10.1109/EMBC.2014.6944123)] [Medline: [25570491](https://pubmed.ncbi.nlm.nih.gov/25570491/)]
18. Maeta K, Nishiyama Y, Fujibayashi K, Gunji T, Sasabe N, Iijima K, et al. Prediction of glucose metabolism disorder risk using a machine learning algorithm: pilot study. *JMIR Diabetes* 2018 Dec 26;3(4):e10212 [FREE Full text] [doi: [10.2196/10212](https://doi.org/10.2196/10212)] [Medline: [30478026](https://pubmed.ncbi.nlm.nih.gov/30478026/)]
19. Anderson JP, Parikh JR, Shenfeld DK, Ivanov V, Marks C, Church BW, et al. Reverse engineering and evaluation of prediction models for progression to type 2 diabetes: an application of machine learning using electronic health records. *J Diabetes Sci Technol* 2015 Dec 20;10(1):6-18 [FREE Full text] [doi: [10.1177/1932296815620200](https://doi.org/10.1177/1932296815620200)] [Medline: [26685993](https://pubmed.ncbi.nlm.nih.gov/26685993/)]
20. Choi SB, Kim WJ, Yoo TK, Park JS, Chung JW, Lee Y, et al. Screening for prediabetes using machine learning models. *Comput Math Methods Med* 2014;2014:618976 [FREE Full text] [doi: [10.1155/2014/618976](https://doi.org/10.1155/2014/618976)] [Medline: [25165484](https://pubmed.ncbi.nlm.nih.gov/25165484/)]
21. Acciaroli G, Sparacino G, Hakaste L, Facchinetti A, Di Nunzio GM, Palombit A, et al. Diabetes and prediabetes classification using glycemic variability indices from continuous glucose monitoring data. *J Diabetes Sci Technol* 2018 Jan;12(1):105-113 [FREE Full text] [doi: [10.1177/1932296817710478](https://doi.org/10.1177/1932296817710478)] [Medline: [28569077](https://pubmed.ncbi.nlm.nih.gov/28569077/)]

22. Shankaracharya, Odedra D, Samanta S, Vidyarthi AS. Computational intelligence-based diagnosis tool for the detection of prediabetes and type 2 diabetes in India. *Rev Diabet Stud* 2012;9(1):55-62 [FREE Full text] [doi: [10.1900/RDS.2012.9.55](https://doi.org/10.1900/RDS.2012.9.55)] [Medline: [22972445](https://pubmed.ncbi.nlm.nih.gov/22972445/)]
23. Wang L, Mu Y, Zhao J, Wang X, Che H. IGRNet: a deep learning model for non-invasive, real-time diagnosis of prediabetes through electrocardiograms. *Sensors (Basel)* 2020 May 30;20(9):1 [FREE Full text] [doi: [10.3390/s20092556](https://doi.org/10.3390/s20092556)] [Medline: [32365875](https://pubmed.ncbi.nlm.nih.gov/32365875/)]
24. Dinh A, Miertschin S, Young A, Mohanty SD. A data-driven approach to predicting diabetes and cardiovascular disease with machine learning. *BMC Med Inform Decis Mak* 2019 Nov 06;19(1):211 [FREE Full text] [doi: [10.1186/s12911-019-0918-5](https://doi.org/10.1186/s12911-019-0918-5)] [Medline: [31694707](https://pubmed.ncbi.nlm.nih.gov/31694707/)]
25. Cahn A, Shoshan A, Sagiv T, Yesharim R, Goshen R, Shalev V, et al. Prediction of progression from pre-diabetes to diabetes: Development and validation of a machine learning model. *Diabetes Metab Res Rev* 2020 Feb;36(2):e3252. [doi: [10.1002/dmrr.3252](https://doi.org/10.1002/dmrr.3252)] [Medline: [31943669](https://pubmed.ncbi.nlm.nih.gov/31943669/)]
26. Garcia-Carretero R, Vigil-Medina L, Mora-Jimenez I, Soguero-Ruiz C, Barquero-Perez O, Ramos-Lopez J. Use of a K-nearest neighbors model to predict the development of type 2 diabetes within 2 years in an obese, hypertensive population. *Med Biol Eng Comput* 2020 May;58(5):991-1002. [doi: [10.1007/s11517-020-02132-w](https://doi.org/10.1007/s11517-020-02132-w)] [Medline: [32100174](https://pubmed.ncbi.nlm.nih.gov/32100174/)]
27. Jin B, Liu R, Hao S, Li Z, Zhu C, Zhou X, et al. Defining and characterizing the critical transition state prior to the type 2 diabetes disease. *PLoS One* 2017;12(7):e0180937 [FREE Full text] [doi: [10.1371/journal.pone.0180937](https://doi.org/10.1371/journal.pone.0180937)] [Medline: [28686739](https://pubmed.ncbi.nlm.nih.gov/28686739/)]
28. Pomares-Quimbaya A, Kreuzthaler M, Schulz S. Current approaches to identify sections within clinical narratives from electronic health records: a systematic review. *BMC Med Res Methodol* 2019 Jul 18;19(1):155 [FREE Full text] [doi: [10.1186/s12874-019-0792-y](https://doi.org/10.1186/s12874-019-0792-y)] [Medline: [31319802](https://pubmed.ncbi.nlm.nih.gov/31319802/)]
29. Honnibal M, Montani I. spaCy 2: Natural language understanding with Bloom embeddings, convolution neural networks and incremental parsing. *spacy.io* 2017:1 [FREE Full text]
30. Loper E, Bird S. NLTK: the natural language toolkit. 2002 Presented at: ACL-02 Workshop on Effective tools and methodologies for teaching natural language processing and computational linguistics; July 7, 2002; Philadelphia, PA URL: <https://doi.org/10.3115/1118108.1118117> [doi: [10.3115/1118108.1118117](https://doi.org/10.3115/1118108.1118117)]
31. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, et al. Scikit-learn: machine learning in Python. *Journal of Machine Learning Research* 2011;12:2825-2830 [FREE Full text] [doi: [10.1145/2786984.2786995](https://doi.org/10.1145/2786984.2786995)]
32. Rouhizadeh M, Jaidka K, Smith L, Schwartz HA, Buffone A, Ungar LH. Identifying locus of control in social media language. 2018 Presented at: Conference on Empirical Methods in Natural Language Processing; October 31 - November 4, 2018; Brussels, Belgium p. 1146-1152 URL: <https://www.aclweb.org/anthology/D18-1145.pdf> [doi: [10.18653/v1/d18-1145](https://doi.org/10.18653/v1/d18-1145)]
33. Park MY, Hastie T. L1 regularization path algorithm for generalized linear models. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)* 2007;69(4):659-677 [FREE Full text] [doi: [10.1111/j.1467-9868.2007.00607.x](https://doi.org/10.1111/j.1467-9868.2007.00607.x)]
34. Hosmer DW, Lemeshow S, Sturdivant RX. *Applied Logistic Regression*, 3rd Edition. Hoboken, NJ: John Wiley & Sons, Ltd; 2013.
35. Cortes C, Vapnik V. Support-vector networks. *Machine Learning* 1995;20:273-297 [FREE Full text] [doi: [10.1007/bf00994018](https://doi.org/10.1007/bf00994018)]
36. Bottou L. Large-Scale Machine Learning with Stochastic Gradient Descent. In: Lechevallier Y, Saporta G, editors. *Proceedings of COMPSTAT'2010*. Heidelberg, Germany: Physica-Verlag HD; 2010:177-186.
37. Safavian S, Landgrebe D. A survey of decision tree classifier methodology. *IEEE Trans. Syst., Man, Cybern* 1991;21(3):660-674 [FREE Full text] [doi: [10.1109/21.97458](https://doi.org/10.1109/21.97458)]
38. Liaw A, Wiener M. Classification and Regression by randomForest. *R News* 2002;2/3:18-22 [FREE Full text]
39. Breiman L. Random forests. *Machine Learning* 2001;45:5-32. [doi: [10.1023/A:1010933404324](https://doi.org/10.1023/A:1010933404324)]
40. Rish I. An Empirical Study of the Naive Bayes Classifier. 2001 Presented at: IJCAI 2001 workshop on empirical methods in artificial intelligence; August 4-6, 2001; Seattle, WA p. 41-46 URL: <https://www.cc.gatech.edu/fac/Charles.Isbell/classes/reading/papers/Rish.pdf>
41. Johnson R, Zhang T. Semi-supervised convolutional neural networks for text categorization via region embedding. *Adv Neural Inf Process Syst* 2015 Dec;28:919-927 [FREE Full text] [Medline: [27087766](https://pubmed.ncbi.nlm.nih.gov/27087766/)]
42. Zhang Y, Wallace B. A sensitivity analysis of (and practitioners' guide to) convolutional neural networks for sentence classification. Cornell University. 2015. URL: <https://arxiv.org/abs/1510.03820> [accessed 2022-01-31]
43. Neumann M, King D, Beltagy I, Ammar W. ScispaCy: Fast and Robust Models for Biomedical Natural Language Processing. Cornell University. 2019. URL: <https://arxiv.org/abs/1902.07669> [accessed 2022-01-31]
44. Honnibal M. Embed, encode, attend, predict: The new deep learning formula for state-of-the-art NLP models. *Explosion AI*. 2016 Nov 9. URL: <https://explosion.ai/blog/deep-learning-formula-nlp> [accessed 2022-01-31]
45. Raschka S. MLxtend: Providing machine learning and data science utilities and extensions to Python's scientific computing stack. *JOSS* 2018 Apr;3(24):638. [doi: [10.21105/joss.00638](https://doi.org/10.21105/joss.00638)]
46. Hu X, Reaven PD, Saremi A, Liu N, Abbasi MA, Liu H, ACT NOW Study Investigators. Machine learning to predict rapid progression of carotid atherosclerosis in patients with impaired glucose tolerance. *EURASIP J Bioinform Syst Biol* 2016 Dec;2016(1):14 [FREE Full text] [doi: [10.1186/s13637-016-0049-6](https://doi.org/10.1186/s13637-016-0049-6)] [Medline: [27642290](https://pubmed.ncbi.nlm.nih.gov/27642290/)]

47. Garcia-Carretero R, Vigil-Medina L, Barquero-Perez O, Ramos-Lopez J. Pulse wave velocity and machine learning to predict cardiovascular outcomes in prediabetic and diabetic populations. *J Med Syst* 2019 Dec 09;44(1):16. [doi: [10.1007/s10916-019-1479-y](https://doi.org/10.1007/s10916-019-1479-y)] [Medline: [31820120](https://pubmed.ncbi.nlm.nih.gov/31820120/)]
48. Zeevi D, Korem T, Zmora N, Israeli D, Rothschild D, Weinberger A, et al. Personalized nutrition by prediction of glycemic responses. *Cell* 2015 Dec 19;163(5):1079-1094 [FREE Full text] [doi: [10.1016/j.cell.2015.11.001](https://doi.org/10.1016/j.cell.2015.11.001)] [Medline: [26590418](https://pubmed.ncbi.nlm.nih.gov/26590418/)]
49. Popp CJ, St-Jules DE, Hu L, Ganguzza L, Illiano P, Curran M, et al. The rationale and design of the personal diet study, a randomized clinical trial evaluating a personalized approach to weight loss in individuals with pre-diabetes and early-stage type 2 diabetes. *Contemp Clin Trials* 2019 Apr;79:80-88. [doi: [10.1016/j.cct.2019.03.001](https://doi.org/10.1016/j.cct.2019.03.001)] [Medline: [30844471](https://pubmed.ncbi.nlm.nih.gov/30844471/)]
50. Liu Y, Wang Y, Ni Y, Cheung CKY, Lam KSL, Wang Y, et al. Gut Microbiome Fermentation Determines the Efficacy of Exercise for Diabetes Prevention. *Cell Metab* 2020 Jan 07;31(1):77-91.e5 [FREE Full text] [doi: [10.1016/j.cmet.2019.11.001](https://doi.org/10.1016/j.cmet.2019.11.001)] [Medline: [31786155](https://pubmed.ncbi.nlm.nih.gov/31786155/)]
51. Fu S, Chen D, He H, Liu S, Moon S, Peterson KJ, et al. Clinical concept extraction: A methodology review. *J Biomed Inform* 2020 Sep;109:103526 [FREE Full text] [doi: [10.1016/j.jbi.2020.103526](https://doi.org/10.1016/j.jbi.2020.103526)] [Medline: [32768446](https://pubmed.ncbi.nlm.nih.gov/32768446/)]

## Abbreviations

**CNN:** convolutional neural network  
**EHR:** electronic health record  
**HbA<sub>1c</sub>:** hemoglobin A<sub>1c</sub>  
**ICD:** International Classification of Diseases  
**IRR:** interrater reliability  
**JHU:** Johns Hopkins University  
**ML:** machine learning  
**NB:** Gaussian naïve bayes  
**NLP:** natural language processing  
**NLTK:** Natural Language Toolkit  
**PCP:** primary care provider  
**PMAP:** Precision Medicine Analytics Platform  
**SGD:** stochastic gradient descent  
**SVM:** support vector machines  
**TF-IDF:** term frequency-inverse document frequency

*Edited by C Lovis; submitted 21.04.21; peer-reviewed by M Peebles, M Burns, M Elbattah, O Serban; comments to author 23.09.21; revised version received 15.11.21; accepted 04.12.21; published 24.02.22.*

*Please cite as:*

*Schwartz JL, Tseng E, Maruthur NM, Rouhizadeh M*

*Identification of Prediabetes Discussions in Unstructured Clinical Documentation: Validation of a Natural Language Processing Algorithm*

*JMIR Med Inform* 2022;10(2):e29803

URL: <https://medinform.jmir.org/2022/2/e29803>

doi: [10.2196/29803](https://doi.org/10.2196/29803)

PMID: [35200154](https://pubmed.ncbi.nlm.nih.gov/35200154/)

©Jessica L Schwartz, Eva Tseng, Nisa M Maruthur, Masoud Rouhizadeh. Originally published in *JMIR Medical Informatics* (<https://medinform.jmir.org>), 24.02.2022. This is an open-access article distributed under the terms of the Creative Commons Attribution License (<https://creativecommons.org/licenses/by/4.0/>), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in *JMIR Medical Informatics*, is properly cited. The complete bibliographic information, a link to the original publication on <https://medinform.jmir.org/>, as well as this copyright and license information must be included.

Original Paper

# Reduction of Platelet Outdating and Shortage by Forecasting Demand With Statistical Learning and Deep Neural Networks: Modeling Study

Maximilian Schilling<sup>1</sup>, BSc; Lennart Rickmann<sup>1</sup>, BSc; Gabriele Hutschenreuter<sup>2</sup>, MD; Cord Spreckelsen<sup>3</sup>, MME, PhD

<sup>1</sup>Institute for Medical Informatics, University Hospital Aachen, RWTH Aachen University, Aachen, Germany

<sup>2</sup>Institute for Transfusion Medicine, University Hospital Aachen, RWTH Aachen University, Aachen, Germany

<sup>3</sup>Institute of Medical Statistics, Computer and Data Sciences, Jena University Hospital, Friedrich Schiller University, Jena, Germany

**Corresponding Author:**

Maximilian Schilling, BSc

Institute for Medical Informatics

University Hospital Aachen

RWTH Aachen University

Pauwelsstraße 30

Aachen, 52074

Germany

Phone: 49 1784599836

Email: [maximilian.schilling@rwth-aachen.de](mailto:maximilian.schilling@rwth-aachen.de)

## Abstract

**Background:** Platelets are a valuable and perishable blood product. Managing platelet inventory is a demanding task because of short shelf lives and high variation in daily platelet use patterns. Predicting platelet demand is a promising step toward avoiding obsolescence and shortages and ensuring optimal care.

**Objective:** The aim of this study is to forecast platelet demand for a given hospital using both a statistical model and a deep neural network. In addition, we aim to calculate the possible reduction in waste and shortage of platelets using said predictions in a retrospective simulation of the platelet inventory.

**Methods:** Predictions of daily platelet demand were made by a least absolute shrinkage and selection operator (LASSO) model and a recurrent neural network (RNN) with long short-term memory (LSTM). Both models used the same set of 81 clinical features. Predictions were passed to a simulation of the blood inventory to calculate the possible reduction in waste and shortage as compared with historical data.

**Results:** From January 1, 2008, to December 31, 2018, the waste and shortage rates for platelets were 10.1% and 6.5%, respectively. In simulations of platelet inventory, waste could be lowered to 4.9% with the LASSO and 5% with the RNN, whereas shortages were 2.1% and 1.7% with the LASSO and RNN, respectively. Daily predictions of platelet demand for the next 2 days had mean absolute percent errors of 25.5% (95% CI 24.6%-26.6%) with the LASSO and 26.3% (95% CI 25.3%-27.4%) with the LSTM ( $P=.01$ ). Predictions for the next 4 days had mean absolute percent errors of 18.1% (95% CI 17.6%-18.6%) with the LASSO and 19.2% (95% CI 18.6%-19.8%) with the LSTM ( $P<.001$ ).

**Conclusions:** Both models allow for predictions of platelet demand with similar and sufficient accuracy to significantly reduce waste and shortage in a retrospective simulation study. The possible improvements in platelet inventory management are roughly equivalent to US \$250,000 per year.

(*JMIR Med Inform* 2022;10(2):e29978) doi:[10.2196/29978](https://doi.org/10.2196/29978)

**KEYWORDS**

platelets; demand forecasting; time series forecasting; blood inventory management; statistical learning; deep learning; LASSO; LSTM



## Introduction

### Background

For blood centers, it is key to keep a balance between shortage and outdated of blood products to secure both cost efficiency and sufficient care for patients. This is especially true for short-lived blood products such as platelets. Forecasting demand has recently gained fresh attention as a way to address the problem, and the rise of *big data* and *artificial intelligence* in recent decades suggests new opportunities in this task [1,2].

Platelet transfusion is an indispensable part of modern medicine [3]. It is used prophylactically to reduce the risk of bleeding or therapeutically to manage active bleeding [3]. Most platelets are transfused to hematology and oncology patients, followed by patients undergoing severe surgical treatment [3-5]. In recent decades, a rise in platelet demand has been reported repeatedly [3,6-8].

As with other blood products, platelets need to be readily available at all times as demand might occur on short notice without obvious foreboding and timely transfusion is often critical [5]. Therefore, most blood centers try to store ample amounts of platelets and other blood products. However, the supply is limited by the number of donations.

Keeping sufficient stock is especially difficult with platelets because of their short shelf life of 5-7 days, including time for preparation and quality control [9]. A large stock may lead to large amounts of wastage because of outdated, whereas a slender stock increases the risk of shortages [10,11]. Platelet outdated rates are the highest of all blood products and are typically reported at 10% to 20% [6,11].

In a recent systematic review, Flint et al [11] provided a detailed overview of existing methods to reduce platelet outdated, one of which was forecasting platelet demand. By forecasting demand, production can be adjusted accordingly to reduce both outdated and shortage. It has been stated that prediction and modeling will have increasingly important roles in managing blood inventory [12]. However, to this day, there are very few scientifically published approaches to forecasting platelet demand [11].

Several authors have investigated different univariate time series models to predict platelet demand, including moving averages, weighted moving averages, exponential smoothing, Winters models, and autoregressive moving averages (ARIMA) [10,13-15]. Fanoodi et al [14] reported improved prediction when using univariate time series modeling by means of an artificial neural network (ANN) compared with an ARIMA model.

More recent studies have included additional clinical data as predictors in multivariate models [1,2,16]. Khaldi et al [16] predicted the monthly demand of platelets, red blood cells, and plasma by means of a multivariate ANN with a total of 10 features, including census data, number of traffic accidents per day, and clinical events such as hemorrhage and deliveries at risk. They reported better prediction accuracy for the ANN compared with a univariate ARIMA model.

Guan et al [1] presented the first *big data* approach to predict platelet demand for the next 3 days and minimize wastage at the Stanford Blood Centre. The authors used 43 features, including hospital census data, complete blood count, day-of-the-week status, and average daily transfusions over the previous 7 days to predict platelet demand [1]. They included the predictions in a linear optimization problem similar to the least absolute shrinkage and selection operator (LASSO) method that also accounted for the structure of the platelet inventory and testing procedure at Stanford Blood Centre to directly minimize wastage [1]. Comparing their findings with retrospective data over 29 consecutive months, Guan et al [1] found that the introduction of such a model in their institution could lower outdated from 10.3% to 3.2% with no shortages.

During the course of this study, Motamedi et al [2] published a study comparing multiple univariate and multivariate models to predict daily platelet demand at Canadian Blood Services: ARIMA, Prophet, LASSO, and a long short-term memory (LSTM) network. They compared the models in terms of prediction errors measured by root mean squared error (RMSE) and mean absolute percent error (MAPE) with 2 and 8 years of training data. The multivariate models (LASSO and LSTM) consistently outperformed univariate time series (ARIMA and Prophet), especially on the shorter training sets. The LASSO performed best, with the LSTM being marginally worse. For the multivariate models, the authors included hospital census data, complete blood count, day-of-the-week status, average transfusions over the previous 7 days, and number of transfusions on the previous day as possible predictors. The features for both the LASSO and the LSTM were selected by the LASSO.

According to the current state of the art, LASSO and LSTM networks seem to be very promising models for the prediction of platelet demand. However, the accuracy of any prediction model may vary between different sites because of the amount and quality of the available data. Furthermore, it is unclear how accurate a prediction needs to be to enable an actual reduction in waste and shortage. This may also vary between sites supposedly because of differences in their respective blood inventories, such as shelf life of platelets, average daily transfusion volume, production and quality control practices, or availability of donations.

### Objective

Therefore, the aims of this study are 2-fold: the first aim is to predict daily platelet demand at the RWTH Aachen University Hospital (UKA) using both a LASSO and an LSTM network. The second aim is to design a simulation model of the blood inventory at UKA, establish an ordering strategy based on the predictions, and quantify possible reductions in waste and shortage rates as compared with retrospective data. To the best of our knowledge, this is the first study to compare these 2 models in terms of both prediction accuracy and possible reduction in waste and shortage rates based on prediction-driven simulations.

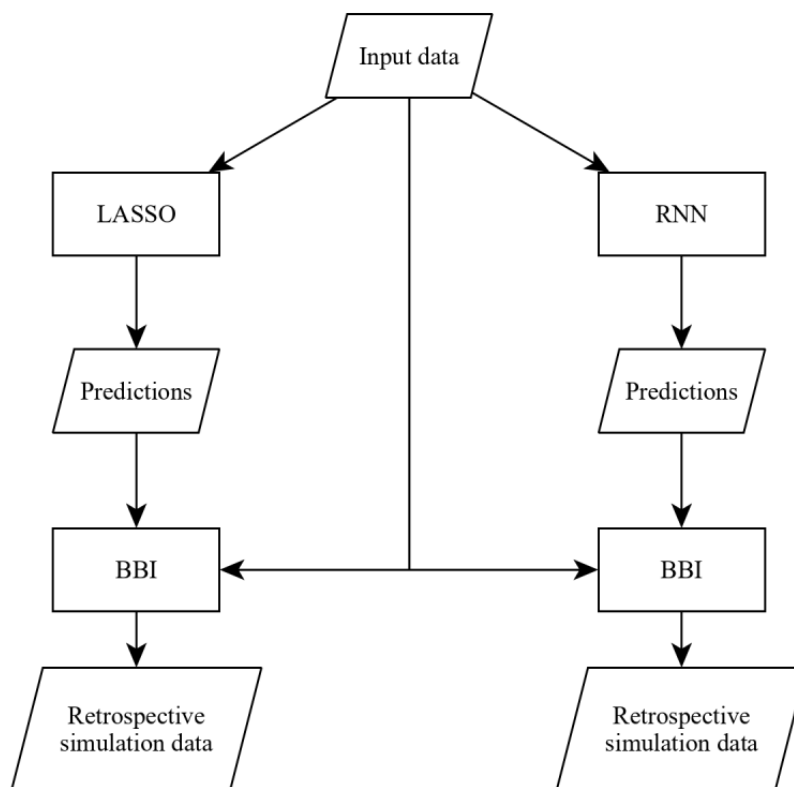
## Methods

### General Approach

According to the aims of this study, our approach was 2-fold (Figure 1). First, we used retrospective data from the UKA electronic health record (EHR) to build 2 separate prediction

models for platelet demand: a LASSO model and a deep learning recurrent neural network (RNN) with an LSTM layer. Second, we designed a simulation model of our blood bank inventory. Combining both parts, forecasts of platelet demand were passed to the blood bank inventory to prematurely adjust platelet production and calculate the resulting expiration and shortage rates in a retrospective simulation study.

**Figure 1.** General approach: input data are fed to 2 separate prediction models—least absolute shrinkage and selection operator and recurrent neural network. Predictions of platelet demand are passed to a simulation model of the blood bank inventory. Possible reductions in waste and shortage rates are calculated in comparison with retrospective data. BBI: blood bank inventory; LASSO: least absolute shrinkage and selection operator; RNN: recurrent neural network.



### Data Acquisition

All data were sourced from the UKA EHR. No personal patient data were used. The local ethics committee approved the data acquisition and analysis (code EK282/19). For the period from January 1, 2008, to December 31, 2018, we obtained data in three categories: (1) platelet ingoings and outgoings as recorded by the transfusion department; (2) census data for all wards, outpatient clinics, and operation rooms; and (3) complete blood count.

### Data Cleaning and Preparation

Data were obtained as a daily time series and aggregated in a single database. Platelet ingoings and outgoings were grouped by source (in-house production and purchase) and disposition (use, waste, sales, and quality control) and documented as platelet units per day. Census data were documented as patients per day grouped by inpatient clinics, outpatient clinics, surgeries, and planned surgeries for the next day and subgrouped by department. Complete blood count data other than platelet count were documented as the number of measurements out of the

norm per day. Platelet count was recorded as the number of measurements per day within specific intervals with regard to platelet transfusion guidelines: <5/nL, 5-10/nL, 10-20/nL, 20-50/nL, 50-70/nL, 70-100/nL, and 100-150/nL [17-20].

Within the UKA EHR, zeroes (eg, no platelets transfused on a given day) are not documented and are represented as missing values. Therefore, we used zeroes to represent the missing values rather than applying imputation. The only exception is census data, where a missing value might indicate that the given department did not exist at that point. Therefore, all departments that did not continually exist throughout the examined 10-year period were excluded. All census data with <400 nonzero values were excluded as it was assumed that these time series did not contain significant information. During the initial inspection of the data, we found that a considerable amount of platelet traffic data was mislabeled in terms of disposition. Over the years, changing collaborations with other clinics and local practices as well as a change in the inventory software have resulted in inconsistent data labeling. A particular problem here was the units that were given to partner clinics but labeled as used in-house rather than sold. Therefore, all platelet traffic data were

systematically verified for correct labeling. Mislabeled data were corrected if possible and excluded otherwise. Less than 1% (305/46,205, 0.66%) of the total transfusion records were excluded because of this problem. The entire data set is provided in [Multimedia Appendix 1](#).

### Included Predictors

All features from the census and complete blood count data with a correlation of  $r^2 \geq 0.2$  to platelet use were included as predictors in the prediction models. Previous studies have shown that platelet transfusion shows a strong pattern of autocorrelation and is dependent on the day of the week [1,10,13]. Therefore, the average number of transfusions per day over the previous 7 days and the day-of-the-week dummy variables were added to the models. Thus, a total of 81 possible features were included for prediction.

### Blood Bank Inventory Model

The UKA transfusion department collects and prepares platelets by apheresis. Registered donors have regular appointments or are called in individually for donation. The entire production chain, including donor activation, platelet preparation, and quality control, takes 2 days (1 day for donor activation and 1 day for preparation and quality control). Donors are only called on Monday through Friday. Therefore, no fresh platelets arrive on Sundays or Mondays. After quality control, platelets have a remaining shelf life of 4 days. In case of slender stock, additional platelets are purchased from other hospitals or local providers such as the local section of the German Red Cross Society. Such an *emergency purchase* is available approximately 2 hours after order. In rare cases, UKA sells platelets to other clinics with a short supply if stock is high. However, as sales occur both very rarely and irregularly, they were not included in the model.

For retrospective simulations of the blood bank inventory, production orders, purchases, discards, and stock are calculated at the end of each day of the observation period using an iterative approach. The stepwise calculation model described below was recalculated for each day of the time series.

As no fresh platelets arrive on Sundays and Mondays, different ordering strategies and prediction intervals for demand are required for different days of the week. Platelets ordered on day  $i$  between Sunday and Wednesday will arrive on day  $i + 2$ . Therefore, these orders need to countervail all platelet outgoings on day  $i + 1$  and  $i + 2$ . Orders made on Thursdays also arrive after 2 days but need to account for the demand of the next 4 days as no orders can be made on Fridays and Saturdays. Considering current stock as well as preceding orders, we established the ordering strategy given in Equation 1, where  $o_i$  is the number of platelets ordered on day  $i$ ,  $\alpha$  is the parameter target value for platelet stock at end of day,  $s_i$  is the current platelet stock at the end of day  $i$ ,  $p_i(2)$  is the predicted demand for days  $i + 1$  and  $i + 2$ ,  $p_i(4)$  is the predicted demand for the next 4 days, and  $o_{i-1}$  is the number of units ordered on day  $i - 1$  as these will arrive on day  $i + 1$ .  $d_w(i)$  represents the weekday status of day  $i$ , with values starting from 0 for Sundays to 6 representing Saturdays.



We established the stepwise calculation model shown in [Figure 2](#) to calculate  $s_i$  as well as other inventory variables. Here,  $r_{x,i}$  represents the remaining units that will be discarded at the end of day  $i + x$ ,  $x$  being the remaining shelf life, with values ranging from 0 to 3 (0 indicating that these units are discarded at the end of that same day).  $u_i$  is the number of platelets actually used on day  $i$ ,  $w_i$  is the number of platelets wasted on day  $i$ , and  $b_i$  is the number of units purchased from other providers on day  $i$ .  $\beta$  and  $\gamma$  are parameters to control for emergency purchases—a purchase is made if stock falls to or below  $\beta$  and, in this case,  $\gamma$  is the target value for stock after emergency purchase.  $t1$ ,  $t2$ ,  $t3$ , and  $t4$  are temporary variables for convenient display. We assume that the oldest platelet units are always the first to be used. The following defaults (indicated as such by the notion *init*) are set each day before moving through the calculation:



After moving through the stepwise calculation,  $s_i$  is calculated to

$$s_i = r_{1,i} + r_{2,i} + r_{a,i} \quad (8)$$

$\alpha$ ,  $\beta$ , and  $\gamma$  are chosen by minimizing the total cost  $c$  as defined by Equation 9 using an exhaustive grid search with a range from 0 to 30 and steps of 1:

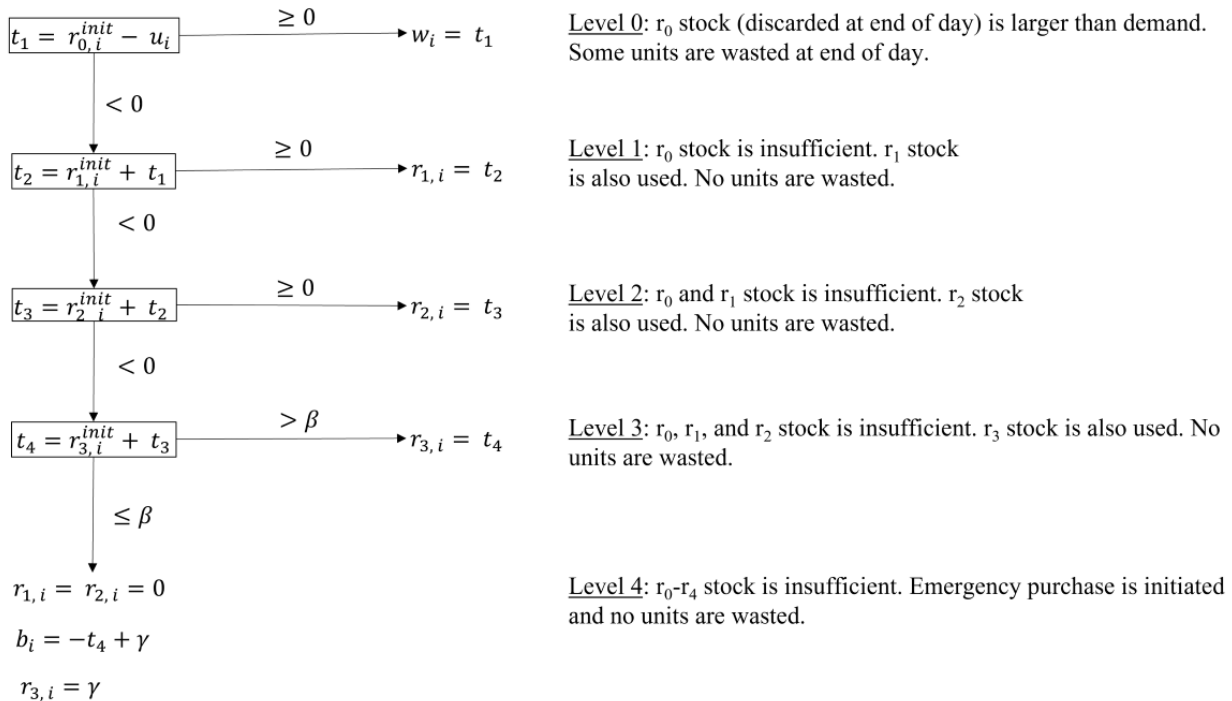


We arrived at this definition because the cost for a single platelet unit is approximately US \$350 when produced locally and planned in advance. Buying platelets in an emergency is more expensive. The actual price varies widely depending on several factors, such as the total amount bought and costs for transportation. On average, the price of a platelet unit bought in an emergency is almost double compared with preplanned production. The weight in Equation 9 was rounded up to also punish the possible delay in transfusion because of transportation time. Note that the blood bank inventory allows for temporarily negative values for stock when moving through the stepwise calculation process given in [Figure 2](#) ( $t1$ ,  $t2$ ,  $t3$ , and  $t4$ ). Therefore, values of 0 for  $\beta$  and  $\gamma$  are possible. In this case, emergency purchase is only initiated when demand exceeds stock ( $\beta=0$ ), and just enough units are bought to satisfy demand, ending that day with empty stock ( $\gamma=0$ ). It is assumed that emergency purchases will always succeed and, therefore, it is simply a matter of buying as many units as required in

circumstances where there is no platelet stock. Consider the following example for  $\beta=\gamma=0$ : stock is 2, and there is an unexpected need for 4 platelet units ( $t_i=4$ ). Emergency purchase is initiated because  $t_i < \beta$ , and 2 units are bought

because  $b_i = -t_i + \gamma = 2 + 0$ . The 2 units from stock and the 2 units from emergency purchase are transfused, and the stock after purchase is 0 ( $\gamma=0$ ).

**Figure 2.** Blood bank inventory stepwise calculation model. For each day of the time series, initial values are set according to Equations 2-7. This stepwise calculation is then carried out and, finally, total stock at end of day is calculated according to Equation 8.



### Prediction Model Setup and Validation Strategy

Standard supervised learning was used to predict platelet demand for the next 2 and 4 days. Predictions were made using rolling-origin-recalibration evaluation as described by Bergmeir and Benítez [21]. First, the models were trained on the first 500 days of the time series. Predictions were made for days 501 to 528. The models were then retrained on the first 528 days, and the next predictions were made for the following 28 days. Both models were retrained in this fashion every 28 days, including recalibration of all hyperparameters. To this end, we also followed the recommendations of Bergmeir and Benítez [21] using 5-fold blocked cross-validation and the augmented Dickey–Fuller unit root test with a trend-corrected regression to check for stationarity in the presence of a trend over time. The interval of 28 days was chosen to account for the weekly seasonality in the data while controlling for the computational expense of repeated retraining [1,10,13]. Mean squared error (MSE) was used as a loss function for the cross-validation. We used the Python 3 language library scikit-learn (Python Software Foundation) to implement this validation strategy [22].

The accuracy of the predictions was measured with RMSE, the Pearson correlation coefficient of the predicted and true values ( $r^2$ ), and MAPE and expressed as mean and 95% CIs. CIs were calculated using bootstrapping [23]. *P* values for the differences in RMSE and MAPE between the models were obtained from the corresponding CI as described by Altman et al [24]. *P* < .05 was defined as statistically significant.

### Statistical Model

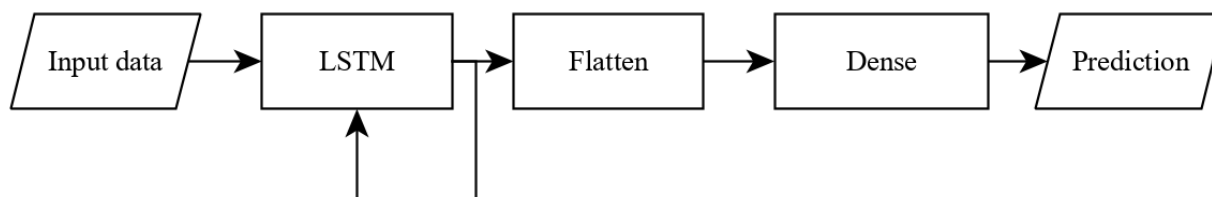
The first model was a LASSO as described by Tibshirani [25]. The LASSO is a shrinkage model for multiple linear regression. Regression coefficients are calculated by minimizing the residual sum of squares with a sparsity penalty given by the L1 norm of the coefficient vector multiplied by a tuning parameter. Owing to the form of the constraint, all coefficients are shrunk toward 0, and some become exactly 0. In this way, the LASSO trades off variance for bias while also performing variable selection and producing interpretable models [25]. As described above, the tuning parameter was chosen via 5-fold blocked cross-validation with MSE as the loss function. We used the Python 3 language library scikit-learn to implement this model [22].

### Deep Learning Model

The second prediction model was an RNN. We used a sequential model from the TensorFlow (Google Brain Team) library (Figure 3) [26]. The first layer was an LSTM as described by Hochreiter and Schmidhuber [27]. An L1–L2 regularizer was combined with a dropout rate to reduce overfitting. The LSTM output was passed to a flatten layer. We treated the prediction of platelet demand as a regression problem and, therefore, used a dense layer with a linear activation function. The dense layer consisted of a single neuron. In preliminary tests on the data, the dropout rate, L1–L2 regularization, batch size, activation function in the flatten layer, and number of units in the LSTM layer were identified as influential hyperparameters. Therefore, they were adjusted during training using a randomized grid

search within the validation strategy described above. All hyperparameters and their search spaces are summarized in [Table 1](#). We used TensorFlow and the Python 3 language library Keras to implement this model [26,28].

**Figure 3.** Architecture of the recurrent neural network used for prediction of platelet demand. Data are first passed to a long short-term memory layer followed by a flatten layer and a dense layer to generate an integer output to our regression problem. LSTM: long short-term memory.



**Table 1.** Hyperparameters of the deep learning model and their respective search space for optimization via randomized grid search.

Parameter	Search space
Batch size	50, 100
LSTM <sup>a</sup> units	10, 50
Dropout rate	0%-50%, steps of 5
L1 regularizer	$10^{-9}$ , $10^{-7}$ , $10^{-5}$ , $10^{-3}$
L2 regularizer	$10^{-9}$ , $10^{-7}$ , $10^{-5}$ , $10^{-3}$
Flatten layer activation function	ReLU <sup>b</sup> , linear

<sup>a</sup>LSTM: long short-term memory.

<sup>b</sup>ReLU: rectified linear unit.

## Results

### Platelet Transfusion, Outdating, and Shortage

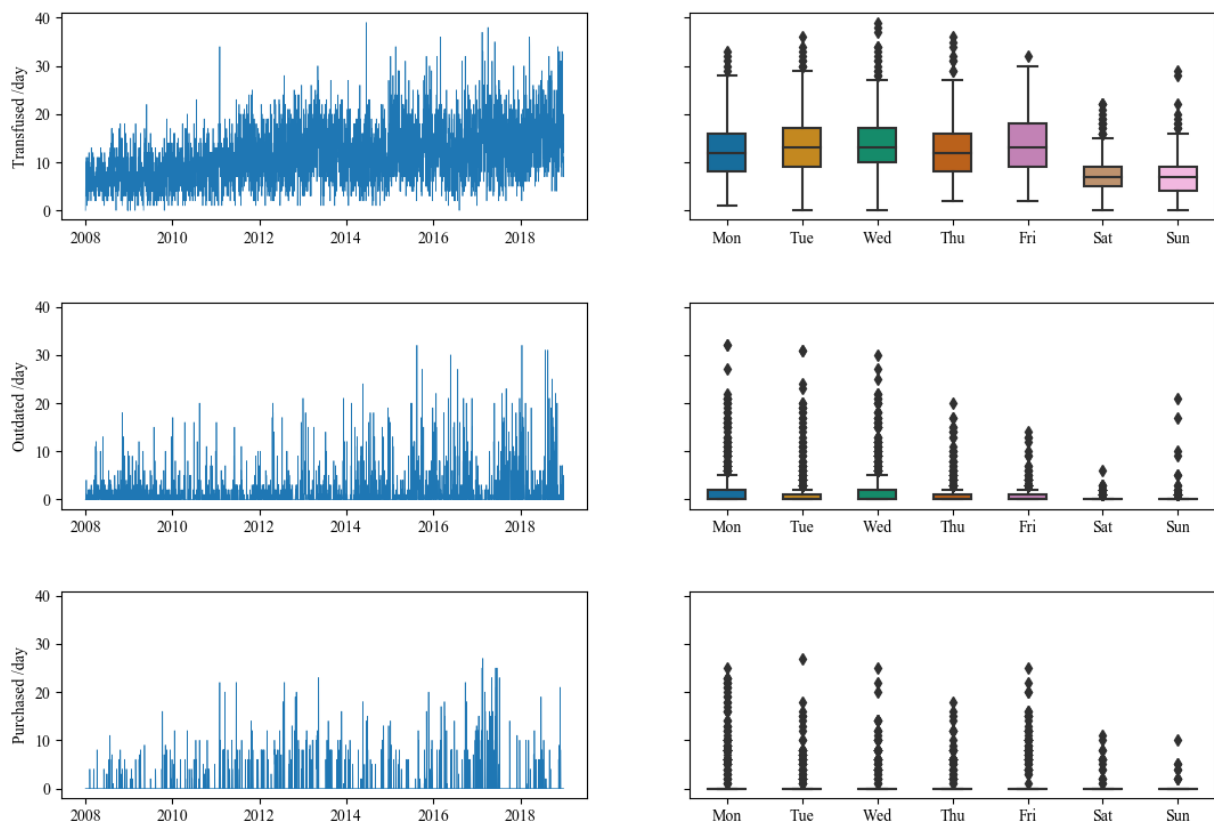
During the observed period, 46,205 platelet units were transfused at UKA. Daily transfusions ranged between 0 and 39 with an average of 11.50 (SD 6.02). Units transfused per year increased from 2566 in 2008 to 5891 in 2018. Daily averages were significantly different for different days of the week as determined by 1-way analysis of variance (ANOVA;  $F_6=187$ ;  $P<.001$ ; [Figure 4](#)). No significant difference was found for month of the year, also by 1-way ANOVA ( $F_{11}=1.56$ ;  $P=.10$ ). More platelets were transfused during the week than on weekends. The time series of daily platelet transfusions was confirmed to be trend-stationary by augmented Dickey–Fuller unit root test with a trend-corrected regression (augmented Dickey–Fuller statistic= $-8.34$ ;  $P<.001$ ).

A total of 4654 platelet units expired during the observed 10 years. The daily average expiration was 1.16 (SD 2.77, range

0-32). Furthermore, 1-way ANOVA showed significant differences in daily platelet expiration across different days of the week ( $F_6=48.9$ ;  $P<.001$ ), with higher values during the week than on weekends ([Figure 4](#)). There was no significant difference across the months of the year ( $F_{11}=1.34$ ;  $P=.20$ ). The expiration rates relative to transfusions were 10.1% and 11% for the entire observed period and the validation period, respectively.

Emergency purchases were made for a total of 2988 units, with a daily mean of 0.74 (SD 2.77, range 0-27). Furthermore, 1-way ANOVA showed significant differences in daily platelet purchases across different days of the week ( $F_6=28.6$ ;  $P<.001$ ; [Figure 4](#)) as well as across the months of the year ( $F_{11}=1.82$ ;  $P=.046$ ). Platelet supply was more often short during the week than during weekends, with most emergency purchases being on Mondays. February and June were the months with the most severe supply shortages. The shortage rates relative to transfusions were 6.47% and 7.05% for the entire observed period and the validation period, respectively.

**Figure 4.** Top to bottom: transfusions, outdating, and emergency purchase of platelet units. Left: daily patterns. Right: averages by day of the week.



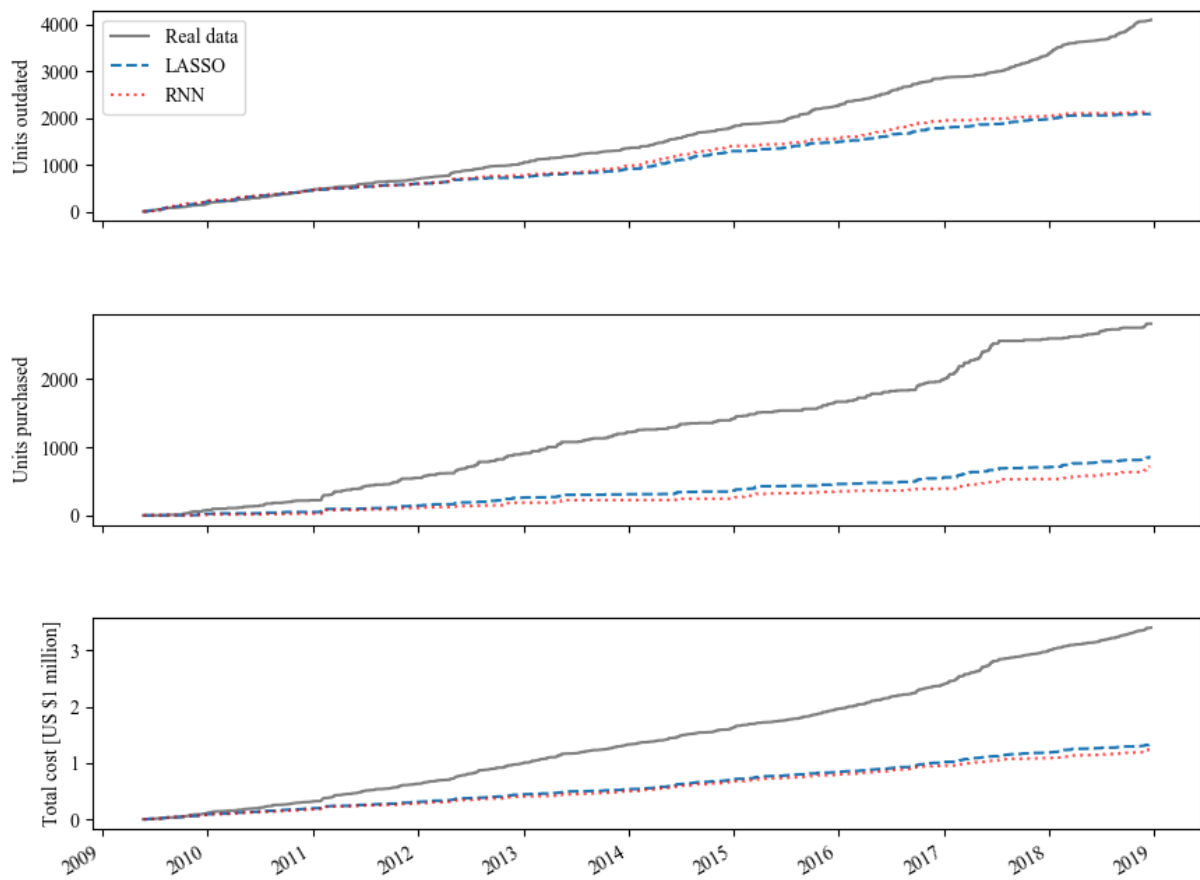
### Blood Bank Inventory Simulation

The retrospective simulations of our blood bank inventory using the above-described blood bank inventory and prediction models yielded the results described in this section. Blood bank inventory simulation was performed separately for predictions made by the LASSO and RNN models. Simulated outdating rates were similar for both prediction methods, whereas purchase and overall cost as defined by Equation 9 were lower with the RNN forecasts. With the LASSO, outdating and shortage were reduced from 11% to 4.93% and from 7.05% to 2.11%, respectively. Using the predictions of the RNN, outdating was reduced to 5%, and shortage fell to 1.68%. These reductions in outdating and shortage are roughly equivalent to savings of US \$250,000 per annum. Simulated total cost was US \$1.33 million with the LASSO and US \$1.241 million with the RNN (Equation

9). Figure 5 shows the cumulative plots for outdating, purchase, and overall cost for both prediction models compared with the real retrospective data.

The target values for platelet stock at the end of each day ( $\alpha$ ) were calculated to be 13 and 14 when using the LASSO and RNN predictions, respectively. The threshold for emergency purchase of platelets ( $\beta$ ) as well as the target value for platelet stock after such purchases ( $\gamma$ ) were 0 for both models. Note that the blood bank inventory allows for temporarily negative values for stock when moving through the stepwise calculation given in Figure 2 ( $t1$ ,  $t2$ ,  $t3$ , and  $t4$ ). Therefore, values of 0 for  $\beta$  and  $\gamma$  mean that emergency purchases are only initiated when demand exceeds current stock ( $\beta=0$ ) and that just enough units are bought to satisfy demand, ending that day with empty stock ( $\gamma=0$ ).

**Figure 5.** Simulated cumulative outdating, purchase, and cost (as defined by Equation 9) compared with retrospective data. LASSO: least absolute shrinkage and selection operator; RNN: recurrent neural network.



**Forecast Accuracy**

Table 2 shows the forecast accuracy for predictions of platelet demand for the next 2 and 4 days measured by RMSE (the square root of the mean square deviation of the predicted values from the true values), the Pearson correlation coefficient of the predicted and true values ( $r^2$ ), and MAPE for both the LASSO and RNN models. The LASSO performed slightly better than

the RNN in terms of these error measures. The differences were statistically significant only for RMSE and MAPE for the 4-day forecast.

Figure 6 shows longitudinal plots of predicted platelet demand alongside the true values for both models and both prediction tasks. Both models trade off variance for bias in their predictions—the RNN more so than the LASSO but with very similar results, as can be seen in Table 2.

**Table 2.** Forecast performance of the least absolute shrinkage and selection operator (LASSO) and recurrent neural network (RNN) for predictions of platelet demand for the next 2 and 4 days.

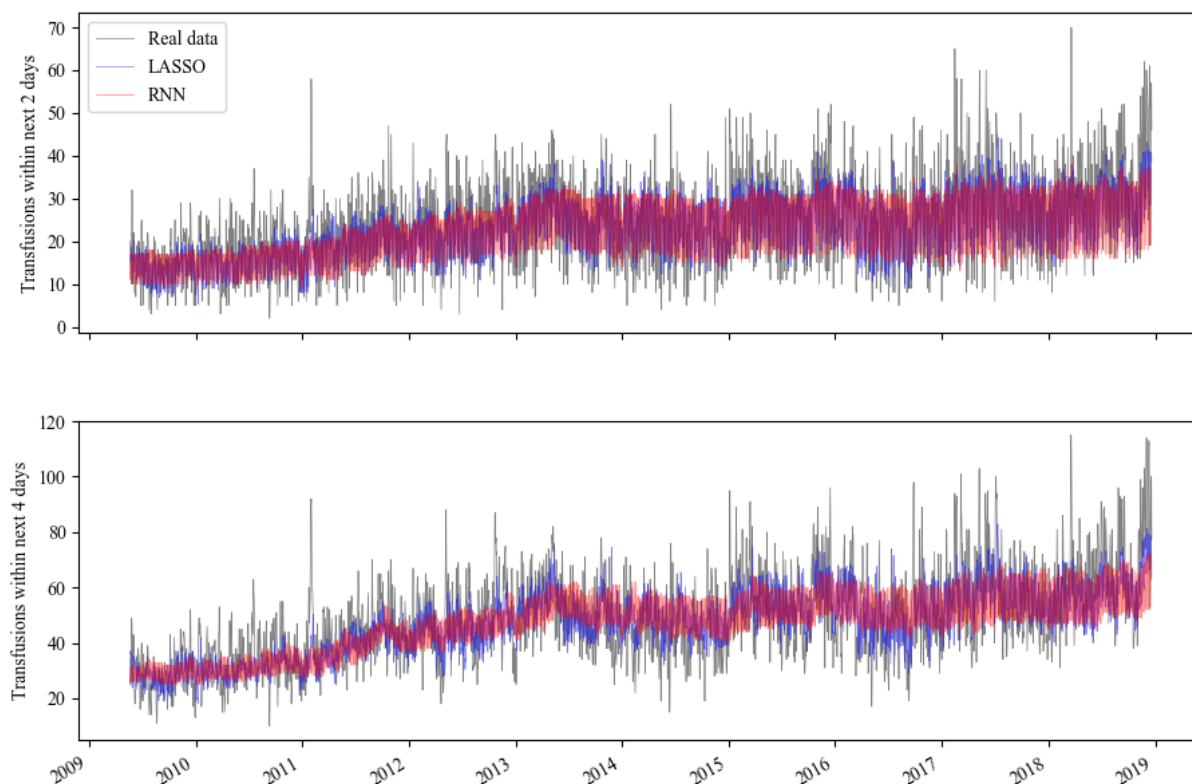
Forecast period and method	RMSE <sup>a</sup> (95% CI)	<i>P</i> value	$r^2$ <sup>b</sup> (95% CI)	<i>P</i> value	MAPE <sup>c</sup> (%; 95% CI)	<i>P</i> value
<b>Next 2 days</b>		.09		.88		.10
LASSO	6.77 (6.57-6.98)		0.73 (0.71-0.74)		25.51 (24.56-26.51)	
RNN	6.94 (6.74-7.15)		0.71 (0.70-0.73)		26.32 (25.33-27.41)	
<b>Next 4 days</b>		<.001		.07		.001
LASSO	10.78 (10.46-11.13)		0.74 (0.72-0.75)		18.11 (17.59-18.61)	
RNN	11.52 (11.17-11.87)		0.69 (0.67-0.71)		19.22 (18.46-19.82)	

<sup>a</sup>RMSE: root mean squared error.

<sup>b</sup>Pearson correlation coefficient of the predictions and the true values.

<sup>c</sup>MAPE: mean absolute percent error.

**Figure 6.** Longitudinal time series plots of demand predictions and real values of platelet demand. LASSO: least absolute shrinkage and selection operator; RNN: recurrent neural network.



### Predictors of Platelet Demand

As described above, the LASSO performs feature selection and produces interpretable models. The most influential predictors of platelet demand for the next 2 and 4 days are listed in [Table 3](#). The strongest predictor in both prediction tasks was the average number of platelet transfusions over the previous 7 days. Other influential predictors were day of the week, number

of platelet counts between 20/nL and 10/nL, patients in the oncology and psychiatry departments, and surgeries planned for the next day in the neurosurgery department. The average number of nonzero predictors over all model iterations was 50.7 (SD 20.409) and 41.8 (SD 14.389) in the 2-day and 4-day forecasts, respectively. Owing to its complex layered structure, the RNN does not provide direct information on the influence of individual predictors.



**Table 3.** Strongest predictors of platelet demand in the least absolute shrinkage and selection operator model. Mean predictor weights over all model iterations.

Forecast and predictor	Predictor weight, mean (SD)
<b>2-day forecast</b>	
PL7 <sup>a</sup>	3.04
Weekday Friday	-2.12
Weekday Thursday	-2.08
I4 <sup>b</sup>	1.54
Weekday Saturday	-1.17
CBC_PL_cont 20-10 <sup>c</sup>	1.17
PP <sup>d</sup>	0.99
OP_P_NC <sup>e</sup>	0.99
<b>4-day forecast</b>	
PL7	1.68
Weekday Saturday	-1.14
Weekday Friday	-1.01
CBC_PL_cont 20-10	0.80
I4	0.64
OP_P_NC	0.61
PP	0.60
OP_P_GG <sup>f</sup>	0.60

<sup>a</sup>PL7: platelet transfusions over previous 7 days.

<sup>b</sup>I4: number of patients in the oncology ward.

<sup>c</sup>CBC\_PL\_cont 10-20: daily number of complete blood count essays with platelet count between >10/nL and ≤20/nL.

<sup>d</sup>PP: number of patients in the psychiatry wards.

<sup>e</sup>OP\_P\_NC: number of planned surgeries for the next day in the neurosurgery department.

<sup>f</sup>OP\_P\_GG: number of planned surgeries for the next day in the vascular surgery department.

## Discussion

### Principal Findings

The results of this study show that it is possible to predict platelet demand at UKA with high accuracy using both approaches investigated: LASSO and RNN with LSTM. These results confirm previous work and, as a particularly relevant aspect, support the generalizability of these models to different sites [1,2].

Furthermore, the simulations of the blood bank inventory suggest that these predictions can be used to reduce waste and shortage of platelets at UKA by a considerable amount. The implementation of such a prediction system at UKA might lead to savings as high as US \$250,000 per year. Although several studies have investigated the prediction of platelet demand, very few have examined the extent to which these predictions can be used to improve inventory management via simulations or field tests [1,2,10,13-16]. To the best of our knowledge, this study is the first to compare LASSO and LSTM models in terms of both prediction accuracy and possible reduction in waste and shortage rates based on prediction-driven simulations.

Both the LASSO and RNNs with LSTM have previously been described as powerful tools for predicting platelet demand [1,2]. Motamedi et al [2] predicted the next-day platelet demand using these models, with very similar results to our study. They reported MAPE values of 28.02% and 28.52% for the LASSO and LSTM, respectively. Guan et al [1] reported possible reduction in platelet outdated from 10.3% to 3.2% with no shortages when using predictions made with the LASSO. However, they did not report the prediction accuracy of their model.

The prediction accuracy of the RNN was marginally inferior to that of the LASSO in our study. This was previously reported by Motamedi et al [2]. However, we argue that the use of deep learning holds great potential not yet fully explored by our project. The most important point is the ability of deep neural networks to take in much more heterogeneous data than a statistical model such as the LASSO [29]. Inclusion of data such as diagnosis and medical history of patients may lead to further refinement of predictions. Despite this potential, the fact that neural networks do not allow for simple interpretation of influential predictors, often referred to as the *Black Box Problem*, is a potential downside of these systems [29-31].

The most influential predictors identified by the LASSO (Table 3) were largely in accordance with previous studies. Previous transfusions and day of the week, the most important predictors in our model, have been described as influential by several authors [1,2,10,13]. In addition, Guan et al [1,2], who also used the LASSO, reported great influence for red cell count and number of patients in the neurosurgery, vascular, and trauma departments. Motamedi et al [2] reported high influence of previous use, day of the week, and abnormal platelet count in their LASSO model. Interestingly, neither of these studies found the number of patients in the hematology and oncology departments to be an influential predictor despite the fact that platelet transfusions are very common in these patients [1-5]. However, this may be due to the intercorrelation effects of the predictors.

As somewhat of an unexpected finding, we observed that the blood bank simulation provided better results in terms of total cost and shortage rates when using RNN predictions, whereas, in accordance with previous results, the predictions made with the LASSO were slightly better in terms of RMSE,  $r^2$ , and MAPE than those of the RNN. Although the differences are small, this indicates that these error measures might not be ideal for the problem. More specifically, the design of the ordering process, as formalized in Equation 1, allows for bias in the predictions to be compensated by the target value for the end-of-day stock ( $\alpha$ ). However, the variance in prediction errors cannot be compensated. Furthermore, because of the platelets' shelf life of 4 days, prediction errors can be (randomly) compensated to some extent by opposing errors within 4 days. Finally, our definition of total cost (Equation 9) punishes shortage more severely than an excess of platelets. These aspects are not adequately represented by error measures such as RMSE,  $r^2$ , or MAPE. In particular, the temporal sequence of errors was not accounted for.

Therefore, we might be missing out on some further reduction in waste and shortage rates by using MSE as a loss function to train the prediction models. Guan et al [1] circumvented this problem by translating demand predictions and modeling of the blood bank inventory into a single optimization problem, thereby using outdated of platelets as a loss function. The problem could also be addressed by replacing MSE as a loss function with error measures that are specifically adapted to the problem at hand. Moreover, this highlights the need for inventory simulation or field tests for any prediction model as the potential to reduce waste and shortage rates is to some extent dependent on the structure and processes of the blood inventory. Further investigation is needed in this area.

### Limitations and Next Steps

With the aforementioned in mind, the modular structure of our system with the prediction models and the blood bank inventory as independent components is a limitation of our study. However, it also has several advantages. First, it reduces the complexity of the overall system. On the one hand, this allows for simple interpretation and comparison of the prediction models. In contrast, it enables the modeling of a very complex blood inventory, incorporating separate predictions for weekdays and weekends as well as emergency purchases while keeping

training times and computational expense manageable as the prediction models do not need to be retrained during the grid search for ideal blood bank inventory parameters. This flexible modular approach will also allow for the addition of further modules, such as a component accounting for blood types in the predictions.

The absence of such a module in our system is another limitation of this study. Although relevant to platelet transfusion, our forecasts do not account for ABO blood types and Rh status [18,32]. There is very limited literature on incorporating blood types in predictions of platelet demand. Critchfield et al [13] used a 7-day moving average of type distribution to account for ABO blood types. Fanoodi et al [14] treated each blood type (ABO and Rh status) as an independent time series for prediction. Although this method is straightforward, it reduces the number of data points available to the prediction models and might lead to reduced prediction accuracy. We suggest the addition of a separate prediction model to our system to forecast blood type distribution of demand. The strong pattern of autocorrelation in platelet demand, supposedly caused by the fact that most patients receive several transfusions over a prolonged period, suggests that the distribution of blood types might also show strong autocorrelation [10,13]. The distribution of blood types in the population could be a further clue to address this problem. Another option is to directly include blood types in a deep learning model based on the RNN presented here as these models are capable of performing complex end-to-end prediction tasks [29].

Although RMSE and MAPE are commonly used in the evaluation of time series forecasts, these error measures might not be the ideal choice here. Further to the potential problems discussed above, their sensitivity to outliers is another limitation [33,34]. As the evaluation of the models did not include testing for significant outliers, they might, if present, cause slight differences in forecast performance between the LASSO and RNN. Therefore, further model refinement should include testing for outliers in the predictions and, if necessary, error measures that are more resilient to outliers, such as MAPE [33].

Although the ordering strategy given by Equation 1 does consider current stock, it neglects the remaining shelf life of units in stock. Adapting orders to the expiry profile of current stock might be beneficial and should be investigated in further studies.

In future applications, the prediction and simulation environment presented here could be extended to other perishable goods whose consumption data show similar characteristics. The following data characteristics may be helpful in generalizing this approach to other problems: (1) the data of platelet demand investigated here are stationary in the presence of a trend, and (2) the data have a strong pattern of autocorrelation with weekly seasonality. From a practical point of view, the short shelf life and high variance of daily demand for platelets are important characteristics that should be considered to identify suitable problems for this approach. Our system could also be used to investigate possible optimization of the blood bank inventory, such as collection of platelets during weekends, by comparing savings in waste and shortage with additional staff costs.

## Conclusions

Both a LASSO model and an RNN with an LSTM layer can predict platelet demand at the UKA with high accuracy. This is in accordance with previous studies and further supports the generalizability of these models to different sites. The retrospective simulations of the blood inventory at the UKA

presented here show that the predictions of both models enable a significant reduction in waste and shortage rates of platelets. Further research is needed to exploit the full potential of deep learning models for the prediction of platelet demand. Furthermore, there is a need for models that take into account ABO blood types in their predictions.

## Authors' Contributions

All authors have agreed on the final version and meet at least one of the following criteria: substantial contributions to conception and design; acquisition of data; or analysis and interpretation of data, drafting of the paper, or revising it critically for important intellectual content.

## Conflicts of Interest

None declared.

## Multimedia Appendix 1

Quantitative data used to construct the figures and tables.

[[XLSX File \(Microsoft Excel File\), 621 KB - medinform\\_v10i2e29978\\_app1.xlsx](#)]

## References

1. Guan L, Tian X, Gombar S, Zemek AJ, Krishnan G, Scott R, et al. Big data modeling to predict platelet usage and minimize wastage in a tertiary care system. *Proc Natl Acad Sci U S A* 2017;114(43):11368-11373 [[FREE Full text](#)] [doi: [10.1073/pnas.1714097114](https://doi.org/10.1073/pnas.1714097114)] [Medline: [29073058](#)]
2. Motamedi M, Li N, Down D, Heddle N. Demand forecasting for platelet usage: from univariate time series to multivariate models. *arXiv* 2021.
3. McCullough J. Overview of platelet transfusion. *Semin Hematol* 2010;47(3):235-242. [doi: [10.1053/j.seminhematol.2010.04.001](https://doi.org/10.1053/j.seminhematol.2010.04.001)] [Medline: [20620434](#)]
4. Wandt H, Schäfer-Eckart K, Greinacher A. Platelet transfusion in hematology, oncology and surgery. *Dtsch Arztebl Int* 2014;111(48):809-815 [[FREE Full text](#)] [doi: [10.3238/arztebl.2014.0809](https://doi.org/10.3238/arztebl.2014.0809)] [Medline: [25512006](#)]
5. Fedele PL, Polizzotto MN, Grigoriadis G, Waters N, Comande M, Borosak M, et al. Profiling clinical platelet and plasma use to inform blood supply and contingency planning: PUPPY, the prospective utilization of platelets and plasma study. *Transfusion* 2016;56(10):2455-2465. [doi: [10.1111/trf.13778](https://doi.org/10.1111/trf.13778)] [Medline: [27600298](#)]
6. Veiho M, Aroviita P, Linna M, Sintonen H, Kekomäki R. Variation of platelet production and discard rates in 17 blood centers representing 10 European countries from 2000 to 2002. *Transfusion* 2006;46(6):991-995. [doi: [10.1111/j.1537-2995.2006.00832.x](https://doi.org/10.1111/j.1537-2995.2006.00832.x)] [Medline: [16734816](#)]
7. Whitaker B, Rajbhandary S, Kleinman S, Harris A, Kamani N. Trends in United States blood collection and transfusion: results from the 2013 AABB blood collection, utilization, and patient blood management survey. *Transfusion* 2016;56(9):2173-2183. [doi: [10.1111/trf.13676](https://doi.org/10.1111/trf.13676)] [Medline: [27301995](#)]
8. Estcourt LJ. Why has demand for platelet components increased? A review. *Transfus Med* 2014;24(5):260-268. [doi: [10.1111/tme.12155](https://doi.org/10.1111/tme.12155)] [Medline: [25327286](#)]
9. Tynngård N. Preparation, storage and quality control of platelet concentrates. *Transfus Apher Sci* 2009;41(2):97-104. [doi: [10.1016/j.transci.2009.07.001](https://doi.org/10.1016/j.transci.2009.07.001)] [Medline: [19699153](#)]
10. McCullough J, Undis J, Allen Jr JW. Platelet production and inventory management. In: Schiffer CA, editor. *Platelet physiology and transfusion: a technical workshop*. Washington: American Association of Blood Banks; 1978:17-37.
11. Flint AW, McQuilten ZK, Irwin G, Rushford K, Haysom HE, Wood EM. Is platelet expiring out of date? A systematic review. *Transfus Med Rev* 2020;34(1):42-50. [doi: [10.1016/j.tmr.2019.08.006](https://doi.org/10.1016/j.tmr.2019.08.006)] [Medline: [31685352](#)]
12. Shah NK. Decision support in transfusion medicine and blood banking. *Clin Lab Med* 2019;39(2):269-279. [doi: [10.1016/j.cll.2019.01.006](https://doi.org/10.1016/j.cll.2019.01.006)] [Medline: [31036280](#)]
13. Critchfield GC, Connelly DP, Ziehwein MS, Olesen LS, Nelson CE, Scott EP. Automatic prediction of platelet utilization by time series analysis in a large tertiary care hospital. *Am J Clin Pathol* 1985;84(5):627-631. [doi: [10.1093/ajcp/84.5.627](https://doi.org/10.1093/ajcp/84.5.627)] [Medline: [4061386](#)]
14. Fanoodi B, Malmir B, Jahantigh FF. Reducing demand uncertainty in the platelet supply chain through artificial neural networks and ARIMA models. *Comput Biol Med* 2019;113:103415. [doi: [10.1016/j.compbiomed.2019.103415](https://doi.org/10.1016/j.compbiomed.2019.103415)] [Medline: [31536834](#)]
15. Kumari DM, Wijayanayake AN. An efficient inventory model to reduce the wastage of blood in the national blood transfusion service. In: 2016 Manufacturing & Industrial Engineering Symposium.: IEEE; 2016 Presented at: MIES'16; October 22, 2016; Colombo p. 1-4. [doi: [10.1109/mies.2016.7779989](https://doi.org/10.1109/mies.2016.7779989)]

16. Khaldi R, El Afia A, Chiheb R, Faizi R. Artificial neural network based approach for blood demand forecasting: Fez transfusion blood center case study. In: Proceedings of the 2nd International Conference on Big Data, Cloud and Applications. New York city: Association for Computing Machinery; 2017 Presented at: BDCA'17; March 29-30, 2017; Tetouan p. 1-6. [doi: [10.1145/3090354.3090415](https://doi.org/10.1145/3090354.3090415)]
17. Kaufman RM, Djulbegovic B, Gernsheimer T, Kleinman S, Tinmouth AT, Capocelli KE, AABB. Platelet transfusion: a clinical practice guideline from the AABB. *Ann Intern Med* 2015;162(3):205-213. [doi: [10.7326/M14-1589](https://doi.org/10.7326/M14-1589)] [Medline: [25383671](https://pubmed.ncbi.nlm.nih.gov/25383671/)]
18. Slichter SJ. Evidence-based platelet transfusion guidelines. *Hematology Am Soc Hematol Educ Program* 2007:172-178. [doi: [10.1182/asheducation-2007.1.172](https://doi.org/10.1182/asheducation-2007.1.172)] [Medline: [18024626](https://pubmed.ncbi.nlm.nih.gov/18024626/)]
19. Bundesärztekammer. Bekanntmachung der querschnitts-leitlinien (BÄK) zur therapie mit blutkomponenten und plasmaderivaten – Gesamtnovelle 2020. *Deutsches Ärzteblatt* 2020;117(40):2.
20. Solves Alcaina P. Platelet transfusion: and update on challenges and outcomes. *J Blood Med* 2020;11:19-26 [FREE Full text] [doi: [10.2147/JBM.S234374](https://doi.org/10.2147/JBM.S234374)] [Medline: [32158298](https://pubmed.ncbi.nlm.nih.gov/32158298/)]
21. Bergmeir C, Benítez JM. On the use of cross-validation for time series predictor evaluation. *Inf Sci* 2012;191:192-213. [doi: [10.1016/j.ins.2011.12.028](https://doi.org/10.1016/j.ins.2011.12.028)]
22. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, et al. Scikit-learn: machine learning in Python. *J Mach Learn Res* 2011;12(85):2825-2830. [doi: [10.3389/fninf.2014.00014](https://doi.org/10.3389/fninf.2014.00014)]
23. James G, Witten D, Hastie T, Tibshirani R. An introduction to statistical learning: with applications in R. New York city: Springer; 2013.
24. Altman DG, Bland JM. How to obtain the P value from a confidence interval. *BMJ* 2011;343:d2304. [doi: [10.1136/bmj.d2304](https://doi.org/10.1136/bmj.d2304)] [Medline: [22803193](https://pubmed.ncbi.nlm.nih.gov/22803193/)]
25. Tibshirani R. Regression shrinkage and selection via the lasso. *J R Stat Soc Series B Stat Methodol* 1996;58(1):267-288. [doi: [10.1111/j.2517-6161.1996.tb02080.x](https://doi.org/10.1111/j.2517-6161.1996.tb02080.x)]
26. Abadi M, Agarwal A, Barham P, Brevdo E, Chen Z, Citro C, et al. TensorFlow: large-scale machine learning on heterogeneous distributed systems. *arXiv* 2016.
27. Hochreiter S, Schmidhuber J. Long short-term memory. *Neural Comput* 1997;9(8):1735-1780. [doi: [10.1162/neco.1997.9.8.1735](https://doi.org/10.1162/neco.1997.9.8.1735)] [Medline: [9377276](https://pubmed.ncbi.nlm.nih.gov/9377276/)]
28. Chollet F. User experience design for APIs. *The Keras Blog*. 2015. URL: <https://blog.keras.io/user-experience-design-for-apis.html> [accessed 2019-11-04]
29. Miotto R, Wang F, Wang S, Jiang X, Dudley JT. Deep learning for healthcare: review, opportunities and challenges. *Brief Bioinform* 2017;19(6):1236-1246. [doi: [10.1093/bib/bbx044](https://doi.org/10.1093/bib/bbx044)] [Medline: [28481991](https://pubmed.ncbi.nlm.nih.gov/28481991/)]
30. Baldi P. Deep learning in biomedical data science. *Annu Rev Biomed Data Sci* 2018;1(1):181-205. [doi: [10.1146/annurev-biodatasci-080917-013343](https://doi.org/10.1146/annurev-biodatasci-080917-013343)]
31. Beam AL, Kohane IS. Big data and machine learning in health care. *JAMA* 2018;319(13):1317-1318. [doi: [10.1001/jama.2017.18391](https://doi.org/10.1001/jama.2017.18391)] [Medline: [29532063](https://pubmed.ncbi.nlm.nih.gov/29532063/)]
32. Bundesärztekammer. Querschnitts-leitlinien zur therapie mit blutkomponenten und plasmaderivaten: mit 19 tabellen. 4. überarb aufl. Köln: Bundesärztekammer; 2009:9783769112696.
33. Armstrong JS. Principles of forecasting: a handbook for researchers and practitioners. Boston: Springer; 2001.
34. Hyndman RJ, Koehler AB. Another look at measures of forecast accuracy. *Int J Forecast* 2006;22(4):679-688. [doi: [10.1016/j.ijforecast.2006.03.001](https://doi.org/10.1016/j.ijforecast.2006.03.001)]

## Abbreviations

- ANN:** artificial neural network
- ANOVA:** analysis of variance
- ARIMA:** autoregressive moving averages
- EHR:** electronic health record
- LASSO:** least absolute shrinkage and selection operator
- LSTM:** long short-term memory
- MAPE:** mean absolute percent error
- MSE:** mean squared error
- RMSE:** root mean squared error
- RNN:** recurrent neural network
- UKA:** RWTH Aachen University Hospital

*Edited by G Eysenbach; submitted 27.04.21; peer-reviewed by G Lim, M Manzanares, C Lynch; comments to author 15.06.21; revised version received 12.09.21; accepted 21.11.21; published 01.02.22.*

*Please cite as:*

*Schilling M, Rickmann L, Hutschenreuter G, Spreckelsen C*

*Reduction of Platelet Outdating and Shortage by Forecasting Demand With Statistical Learning and Deep Neural Networks: Modeling Study*

*JMIR Med Inform 2022;10(2):e29978*

*URL: <https://medinform.jmir.org/2022/2/e29978>*

*doi: [10.2196/29978](https://doi.org/10.2196/29978)*

*PMID: [35103612](https://pubmed.ncbi.nlm.nih.gov/35103612/)*

©Maximilian Schilling, Lennart Rickmann, Gabriele Hutschenreuter, Cord Spreckelsen. Originally published in JMIR Medical Informatics (<https://medinform.jmir.org>), 01.02.2022. This is an open-access article distributed under the terms of the Creative Commons Attribution License (<https://creativecommons.org/licenses/by/4.0/>), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Medical Informatics, is properly cited. The complete bibliographic information, a link to the original publication on <https://medinform.jmir.org/>, as well as this copyright and license information must be included.

Original Paper

# Comparison of the Erectile Dysfunction Drugs Sildenafil and Tadalafil Using Patient Medication Reviews: Topic Modeling Study

Maryanne Kim<sup>1,2</sup>, MS, PharmD; Youran Noh<sup>1,2</sup>, MS; Akihiko Yamada<sup>1</sup>, PhD; Song Hee Hong<sup>1,2</sup>, PhD

<sup>1</sup>College of Pharmacy, Seoul National University, Seoul, Republic of Korea

<sup>2</sup>Research Institute of Pharmaceutical Sciences, Seoul National University, Seoul, Republic of Korea

**Corresponding Author:**

Song Hee Hong, PhD

College of Pharmacy

Seoul National University

Suite 20-322

1 Gwanak-ro, Gwanak-gu

Seoul, 08826

Republic of Korea

Phone: 82 2 880 1547

Fax: 82 2 877 7589

Email: [songhhong@snu.ac.kr](mailto:songhhong@snu.ac.kr)

## Abstract

**Background:** Topic modeling of patient medication reviews of erectile dysfunction (ED) drugs can help identify patient preferences regarding ED treatment options. The identification of a set of topics important to the patient from social network service drug reviews would inform the design of patient-centered medication counseling.

**Objective:** This study aimed to (1) identify the distinctive topics from patient medication reviews unique to tadalafil versus sildenafil; (2) determine if the primary topics are distributed differently for each drug and for each patient characteristic (age and time on ED drug therapy); and (3) test if the primary topics affect satisfaction with ED drug therapy controlling for patient characteristics.

**Methods:** Data were collected from the patient medication reviews of sildenafil and tadalafil posted on WebMD and Ask a Patient. The latent Dirichlet allocation method of natural language processing was used to identify 5 distinctive topics from the patient medication reviews on each drug. Analysis of variance and a 2-sample *t* test were conducted to compare the topic distribution and assess whether patient satisfaction varies with the primary topics, age, and time on medication for each ED drug. Statistical significance was tested at an alpha of .05.

**Results:** The patient medication reviews of sildenafil (N=463) had 2 topics on treatment benefit and 1 each on medication safety, marketing claim, and treatment comparison, while the patient medication reviews of tadalafil (N=919) had 2 topics on medication safety and 1 each on the remaining subjects. Sildenafil's reviewers quite frequently (94/463, 20.4%) mentioned *erection sustainability* as their primary topic, whereas tadalafil's reviewers were more concerned about *severe medication safety*. Those who mentioned *erection sustainability* as their primary topic were quite satisfied with their treatment as opposed to those who mentioned *severe medication safety* as their primary topic (score 3.85 vs 2.44). The discovered topics reflected the marketing claims of *blue magic* and *amber romance* for sildenafil and tadalafil, respectively. The topic of *blue magic* was preferred among younger patients, while the topic of *amber romance* was preferred among older patients. The topic *alternative choices*, which appeared for both the ED drugs, reflected patient interest in the comparative effectiveness and price outside the drug labeling information.

**Conclusions:** The patient medication reviews of ED drugs reflect patient preferences regarding drug labeling information, marketing claims, and alternative treatment choices. The patient preferences concerning ED treatment attributes inform the design of patient-centered communication for improved ED drug therapy.

(JMIR Med Inform 2022;10(2):e32689) doi:[10.2196/32689](https://doi.org/10.2196/32689)

## KEYWORDS

topic modeling; patient preference; patient-centered communication; erectile dysfunction; PDE5 inhibitor; phosphodiesterase type 5 inhibitor

## Introduction

Topic modeling has been used frequently in various health care fields, including clinical research and health communication, for uncovering themes hidden in natural languages. For example, topic modeling has been used to characterize people's opinions about vaccines communicated on Twitter [1], to predict clinical outcomes using notes on electronic health records [1,2], and to identify patients' medical conditions from referral letters [3]. Topic modeling has also been applied in pharmacovigilance to identify drugs with similar safety concerns and therapeutic uses based on the Food and Drug Administration (FDA) drug labeling information [4,5].

Recently, topic modeling on data collected via social network services (SNSs), such as Twitter, and web portals is widely used for the survey of public perceptions and attitudes toward the COVID-19 outbreak [6,7], containment strategies [8,9], treatment interventions [6], and vaccines [10,11]. Topic modeling on SNS data is useful for examining issues that change quickly over time [12]. Topic modeling is especially useful for studying private and sensitive issues such as abortion [13], domestic violence [14], and bullies [15]. On SNSs, people freely reveal their honest attitudes and opinions, while being reluctant to do so on formal surveys when their attitudes and opinions contradict social desirability [16,17]. In fact, a recent study reported that adults in mainland China actively search the internet for information on premature ejaculation [18].

The drug reviews on SNSs can be regarded as patient-reported outcomes (PROs) conveyed in natural language. Directly coming from patients without clinician filtering or interpretation, drug reviews represent the treatment effectiveness and medication safety experienced by individual patients [19-21]. Drug reviews therefore likely contain the labeling information approved by the regulatory agency. They also likely include the marketing claims meticulously chosen by sellers to emphasize the treatment benefits. Furthermore, drug reviews may comprise any other information important to the patient whose real-world experience may well be different from that in the trial setting [21,22]. Therefore, the identification of a set of topics important to the patient from SNS drug reviews would inform the design of patient-centered medication counseling, comparative effectiveness research, pharmacovigilance, and marketing.

The 2 erectile dysfunction (ED) drugs sildenafil and tadalafil have been competing as phosphodiesterase type 5 (PDE5) inhibitors for more than 10 years. However, very little is known about what really concerns the patients who take the medication. This study aimed to identify the topics mentioned in SNS drug reviews by patients who had taken an ED drug (sildenafil vs tadalafil). The study's specific aims were to determine if (1) the topics identified for each ED drug reflect drug labeling information, marketing claims, and other patient concerns; (2) the distribution of primary topics varies with patient characteristics (patient age and time on ED drug therapy); and

(3) the satisfaction with ED drug therapy depends on the primary topics controlling for patient characteristics.

## Methods

### Study Design and Settings

Data were collected from the patient reviews on WebMD [23] and Ask a Patient [24] in the United States. Both WebMD and Ask a Patient are health social media that allow patients to browse patient reviews of prescription drugs based on their medication experience and post their own reviews. Patient reviews on WebMD consist of 4 fields. Reviewers can choose a reason for taking the drug, among several possible reasons given by WebMD. There is an open-ended comments section where reviewers can share their treatment experiences, including benefits, medication safety, and how or whether it worked. They can also give their information (optional), such as age and time on medication, by choosing from a list of options. Finally, patients can rate their drug experience in terms of effectiveness, ease of use, and overall satisfaction. The ratings are based on a 5-point Likert scale from 1 (least satisfied) to 5 (most satisfied). Patient reviews on Ask-a-Patient have 8 fields, namely overall satisfaction rating, reason for taking the drug, side effects, comments, gender, age, duration/dosage, and date. Most of the fields are filled manually by the reviewers. They can also rate their treatment based on a 5-point Likert scale provided by the website. To align with the reviews on Ask a Patient, only the overall satisfaction drug rating was selected from WebMD ratings.

### Data Collection

The drug reviews posted prior to July 1, 2019, were collected for both ED drugs. Among the collected reviews on WebMD, posts without any comments were removed. Since Ask a Patient reviews have a separated comments section regarding side effects, the posts without any comments were removed. To exclude spam, we identified and removed the reviews containing "http," ".com," or ".www." Reviews by those under the age of 19 years or without age information were also excluded. Reviews by females were not excluded since they may have been written by caregivers or partners who can represent the user's experience. The reviews were freely available to all web users and did not include private identifiable data. According to the guidelines, in most cases, research involving such reviews is classified as nonhuman research.

### Text Processing

Preprocessing involving tokenization, stop words, stemming, and completion was used to process the content of patient reviews using the R package. A corpus created based on a list of words was then cleaned by removing punctuations, numbers, extra white spaces, and irrelevant words. Typographical errors were also corrected to prevent the errors from being processed as separate words. Bigrams consisting of 2 words frequently appearing together such as "erectile dysfunction" and "side

effect” were treated as unigrams before stemming. Stemming was done to reduce inflected words to their word stem. The stemmed words were then replaced with the most prevalently appearing words from the reviews. Finally, a document-term matrix consisting of words along with their frequencies was constructed.

### Topic Modeling

The latent Dirichlet allocation (LDA) method of natural language processing (NLP) was used to discover hidden topics from each set of patient reviews [25]. The algorithm treated each review as a mixture of several topics and each topic as a distribution of words. To identify the correct weights between these matrices, Gibbs sampling was used. For the LDA topic modeling, the number of topics, “5,” was given to each drug. The optimality of the 5 topics was determined based on a density-based method and visualization to find distinctive and independent topics [26-28]. The LDA packages of open-source R were used as the analysis tool. The primary topic was defined as the topic most frequently mentioned in each review [19].

### Drug Labeling Information and Marketing Claims

Drug labeling information for sildenafil and tadalafil was accessed from the drug database of the FDA [29]. The labeling information comprises efficacy, safety, and dosing schedules. Efficacy is measured based on the PROs on erection strength, duration, etc. The evidence on safety documents headaches, nasal congestion, back pain, and muscle pain. Sildenafil has additional safety concerns pertaining to abnormal vision and rash, while tadalafil has an additional safety concern related to pain in the limbs. The approved dosing schedules specify that sildenafil acts for 4 hours as opposed to tadalafil that has an effect up to 36 hours without being affected by food and liquid intake.

With regard to marketing claims, sildenafil was marketed as the “blue pill” or “blue diamond,” with sports stars of the time promoting the slogan “Get back to Mischief.” At the same time, sildenafil was promoted as a recreational aid to expand the consumer base rather than as a medical treatment [30,31]. On the other hand, tadalafil was publicized as fostering a romantic relationship. It was marketed as a drug that makes you ready

whenever you feel the urge to make love, especially during weekends, guaranteeing 36 hours of confidence. Furthermore, it was advertised that users can drink and eat while being on the drug [32].

### Statistical Analysis

The frequency of each topic was computed for each review and then summed for all reviews. The Fisher exact test was used to compare the topic distribution between sildenafil and tadalafil. The 2-sample *t* test was performed to test whether the patient medication ratings varied with primary topics, age, and the time on medication between the drugs. Analysis of variance was used to compare the ratings of the medication for each primary topic by age and time on medication. Statistical significance was tested at an alpha of .05.

## Results

### Description of Patient Medication Reviews

The total number of patient reviews posted on Ask a Patient and WebMD was 1567 (547 for sildenafil and 1020 for tadalafil). The number reduced to 1382 (463 for sildenafil and 919 for tadalafil) when ineligible reviews (those without comments, commercial posts, and reviews by those below 19 years of age) were excluded (Table 1). Most of the reviews were from the age group of 45-64 years (ie, 257/463, 55.5% for sildenafil and 559/919, 60.8% for tadalafil). They were mostly written by patients who used the medication for less than a month (163/463, 35.2% for sildenafil and 448/919, 48.7% for tadalafil). Additionally, most reviews were posted by the patients themselves (189/203, 93.1% for sildenafil and 311/343, 90.7% for tadalafil), while a few (less than 4%) were posted by caregivers.

Among the reasons for taking the drug, “Inability to have an erection” was the most common one for both drugs according to WebMD (166/203, 81.8% for sildenafil and 253/343, 73.8% for tadalafil). However, the reason for taking the drug is not clearly distinguished on Ask a Patient since the reviewer has to write manually rather than choose from a list. The reviewers were dominantly males (more than 94% for both drugs); female reviewers were either caregivers or partners of the drug users.



**Table 1.** Characteristics of patient medication reviews.

Demographic	Sildenafil (N=463), n (%)	Tadalafil (N=919), n (%)
<b>Gender</b>		
Male	445 (96.1)	869 (94.6)
Female	7 (1.5)	16 (1.7)
Not available	11 (2.4)	34 (3.7)
<b>Age (years)</b>		
19-44	124 (26.8)	225 (24.5)
45-64	257 (55.5)	559 (60.8)
≥65	82 (17.7)	135 (14.7)
<b>Time on medication</b>		
<1 month	163 (35.2)	448 (48.7)
1 month to <1 year	141 (30.5)	240 (26.1)
≥1 year	150 (32.4)	187 (20.3)
Not available	9 (1.9)	44 (4.8)
<b>Reasons for taking medications (WebMD)<sup>a</sup></b>		
Inability to have an erection	166 (81.8)	253 (73.8)
Increased pressure of pulmonary circulation	5 (2.4)	5 (1.5)
Pulmonary arterial hypertension	2 (1.0)	1 (0.3)
Enlarged prostate	__ <sup>b</sup>	22 (6.4)
Enlarged prostate with urination problems	__ <sup>b</sup>	10 (2.9)
Other	30 (14.8)	52 (15.2)
<b>Reviewer type (WebMD)<sup>a</sup></b>		
Caregiver	3 (1.5)	13 (3.8)
Patient	189 (93.1)	311 (90.7)
Not available	11 (5.4)	19 (5.5)
<b>Year</b>		
2001-2004	21 (4.5)	__ <sup>b</sup>
2005-2009	201 (43.4)	343 (37.3)
2010-2014	179 (38.7)	424 (46.1)
2015-2019	62 (13.4)	152 (16.5)

<sup>a</sup>Only the reviews posted on WebMD have this information.

<sup>b</sup>Not available.

### Identification of Distinctive Topics of Patient Medication Experiences

The number of distinctive topics identified was 5 for each ED drug (Textbox 1). The identified topics were subjectively named based on the top 30 most frequently appearing words. They represented treatment benefits such as *sexual performance* for tadalafil and sildenafil, and *erection sustainability* for sildenafil. They also reflected marketing tags such as *blue magic* for sildenafil and *amber romance* for tadalafil. As for medication safety, sildenafil had a topic named *medication safety* for which events are known to be typical of PDE5 inhibitors, while

tadalafil had 2 topics named *mild medication safety* and *serious medication safety*. *Alternative choices*, which is the only topic representing patient concern outside drug labeling information, was identified in both ED drugs.

In addition to the topic of *alternative choices*, *sexual performance* was demonstrated for both ED drugs. As for the topics on medication safety, they were identified in both ED drugs but with different grades, that is, typical safety for sildenafil, and serious and mild safeties for tadalafil. *Erection sustainability* was only observed with sildenafil. As for the topics related to the marketing claims, *blue magic* and *amber romance* were identified accordingly for the respective drugs.

**Textbox 1.** List of 5 topics and their member words (top 30 frequently appearing words) identified for each drug.

<p><b>Sildenafil (N=461)</b></p> <p><b>- Sexual performance (n=102, 22.1%)</b></p> <p>Words: erect, get, hard, wife, can, result, good, orgasm, still, experience, longer, best, without, cut, penile, need, increase, notice, enough, stay, stimulated, keep, taken, rock, ejaculate, flush, morning, since, position, and sexual</p> <p><b>- Erection sustainability (n=94, 20.4%)</b></p> <p>Words: time, last, sex, pill, first, great, long, get, start, doctor, make, got, back, month, little, medical, life, week, morning, love, problem, always, recommend, couple, worth, ever, help, made, stop, and way</p> <p><b>- Medication safety (n=104, 22.6%)</b></p> <p>Words: headache, drug, flush, work, feel, nose, face, eye, stuffiness, slight, mild, red, light, vision, blue, sometime, think, congested, side effect, nasal, pressure, facial, less, nothing, stuff, head, say, drink, seems, and well</p> <p><b>- Alternative choices (n=71, 15.4%)</b></p> <p>Words: Viagra, work, use, trial, side effect, year, problem, Cialis, cause, give, erectile dysfunction, well, intercourse, help, pain, take, blood, due, another, full, generic, perform, sex, year old, high, maintain, wait, find, gave, and never</p> <p><b>- Blue magic (n=90, 19.5%)</b></p> <p>Words: take, hour, effect, like, day, dose, took, much, half, heart, minute, stomach, felt, night, later, know, want, within, start, better, min, bad, beat, med, several, tablet, usual, around, away, and rapid</p>
<p><b>Tadalafil (N=915)</b></p> <p><b>- Sexual performance (n=166, 18.1%)</b></p> <p>Words: erect, get, hard, sex, wife, can, like, problem, long, need, night, start, enough, able, better, longer, orgasm, sometime, good, keep, life, several, occasion, lot, penile, love, minute, full, perform, and quit</p> <p><b>- Serious medication safety (n=244, 26.7%)</b></p> <p>Words: pain, day, back, leg, bad, lower, severe, ache, sleep, never, worth, muscle, cramp, symptom, stop, still, extreme, away, due, upper, right, thigh, terrible, though, hip, way, ever, like, walk, and neck</p> <p><b>- Mild medication safety (n=181, 19.8%)</b></p> <p>Words: take, side effect, drug, effect, dose, experience, erectile dysfunction, flush, cause, start, eye, great, help, much, mild, bodies, dosage, side, blood, issue, however, increase, read, recommend, taken, heart, wonder, continuation, face, and sore</p> <p><b>- Alternative choices (n=182, 19.9%)</b></p> <p>Words: Cialis, work, use, trial, year, Viagra, week, well, doctor, month, make, good, result, medical, best, Levitra, see, couple, nothing, gave, vision, anyone, later, sexual, since, always, guy, happen, per, and generic</p> <p><b>- Amber romance (n=142, 15.5%)</b></p> <p>Words: time, hour, last, headache, took, pill, first, feel, morning, half, tablet, night, like, great, got, slight, heartburn, nose, thing, notice, felt, stuffiness, give, still, went, much, three, within, think, and weekend</p>

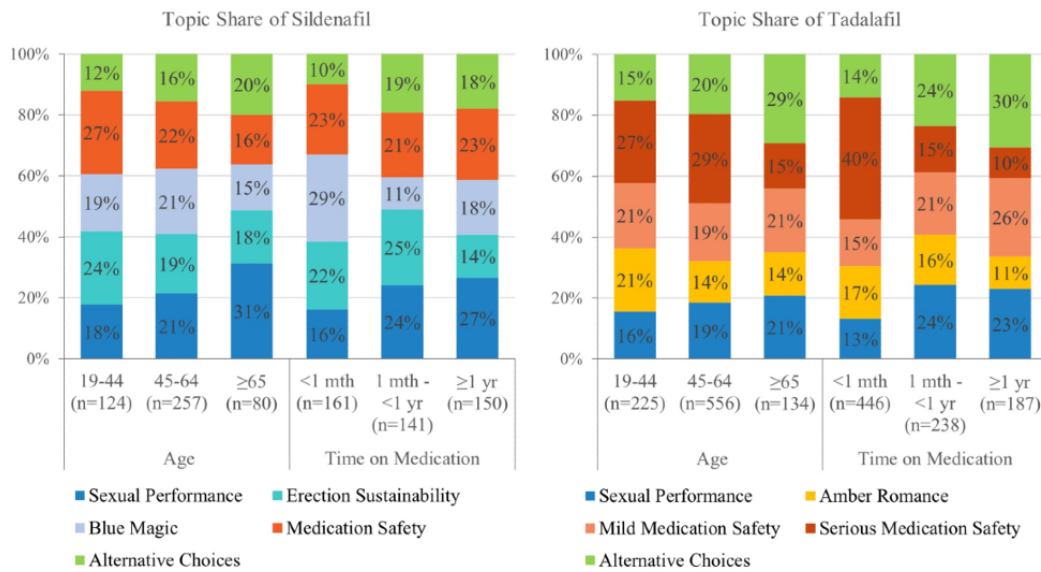
### Primary Topics by Age and Time on Medication

The topic identified would be primary if it occurs most frequently in a patient medication review. The shares of primary topics varied with patient characteristics (Figure 1). The oldest reviewers of sildenafil most frequently mentioned *sexual performance* as the primary topic, followed by *alternative choices*. However, the oldest tadalafil reviewers most frequently mentioned *alternative choices*, followed by *mild medication safety*. As the reviewers' age increased, *sexual performance* and *alternative choices* were more likely the primary topics, and *medication safety* was less likely the primary topic. *Medication safety* and *erection sustainability* were most likely

the primary topics among youngest sildenafil reviewers, while *serious medication safety* and *amber romance* were most likely the primary topics among youngest tadalafil reviewers.

The most frequently occurring topic did vary with time on the ED drug. Reviewers who experienced the longest time on sildenafil more likely mentioned *sexual performance*. However, those who experienced the shortest time on sildenafil more likely mentioned *medication safety*. The patient reviewers with the shortest time on tadalafil also most likely mentioned *medication safety*, specifically *serious medication safety*. In contrast, those with the longest time on tadalafil least likely mentioned *serious medication safety*.

**Figure 1.** Topic distribution of erectile dysfunction therapy by age and time on medication. mth: month; yr: year.



**Drug Ratings by Primary Topic, Age, and Time on Medication**

Drug ratings depended on what topic the reviewers would most likely mention ( $P=.02$  for sildenafil and  $P<.001$  for tadalafil). Those who mentioned *sexual performance* or *erection sustainability* as their primary topic gave higher ratings than those who mentioned *medication safety* as their primary topic. The dependency of the drug ratings on each primary topic further varied with age as well as time on medication (Tables 2 and 3). Among the sildenafil reviewers, the primary topic of *erection sustainability* had the largest variation in drug ratings across different ages (4.37 for the youngest group compared with 2.86 for the oldest group), followed by the primary topic of *alternative choices* (4.33 for the youngest group compared with 2.94 for the oldest group). The least variation in drug rating across different ages was observed with the primary topic of *blue magic*, which indicates that those reviewers mentioning the primary topic of *blue magic* gave consistent drug ratings regardless of age. Among the sildenafil reviewers, those with the primary topic of *medication safety* had the reverse order of drug rating across ages, with the youngest group giving the lowest rating of 2.56. However, among the tadalafil reviewers, age variation was not apparent, except for the primary topic of *sexual performance*. The youngest group with the primary topic

of *sexual performance* gave a rating of 4.17, while the oldest group gave a rating of 3.29. Those with the primary topic of *serious medication safety* reported a drug rating of 2.5 or less across age groups, whereas those with the primary topic of *mild medication safety* reported a drug rating of 3.23-3.64.

When comparing the drug therapy, medication reviewers rated sildenafil 0.29 points ( $P<.001$ ) higher than tadalafil (Figure 2). Medication reviewers who mentioned topics about treatment benefits, such as *sexual performance* and *erection sustainability*, as their primary topics rated sildenafil lower than tadalafil (3.90 vs 4.14). However, reviewers who mentioned *medication safety* as their primary topic gave the lowest drug rating to each drug, and tadalafil received a lower rating compared with sildenafil (3.30 vs 2.90). Among those who mentioned marketing claims as their primary topic, tadalafil was rated better than sildenafil (3.58 vs 3.81).

When the drug ratings were examined by age and time on medication, the oldest reviewers gave tadalafil a slightly better rating, while younger reviewers gave sildenafil a better rating. Patients aged between 19 and 44 years gave about 0.42 points more for sildenafil than for tadalafil. A longer time on medication was associated with a better rating for the ED drug regardless of drug therapy.

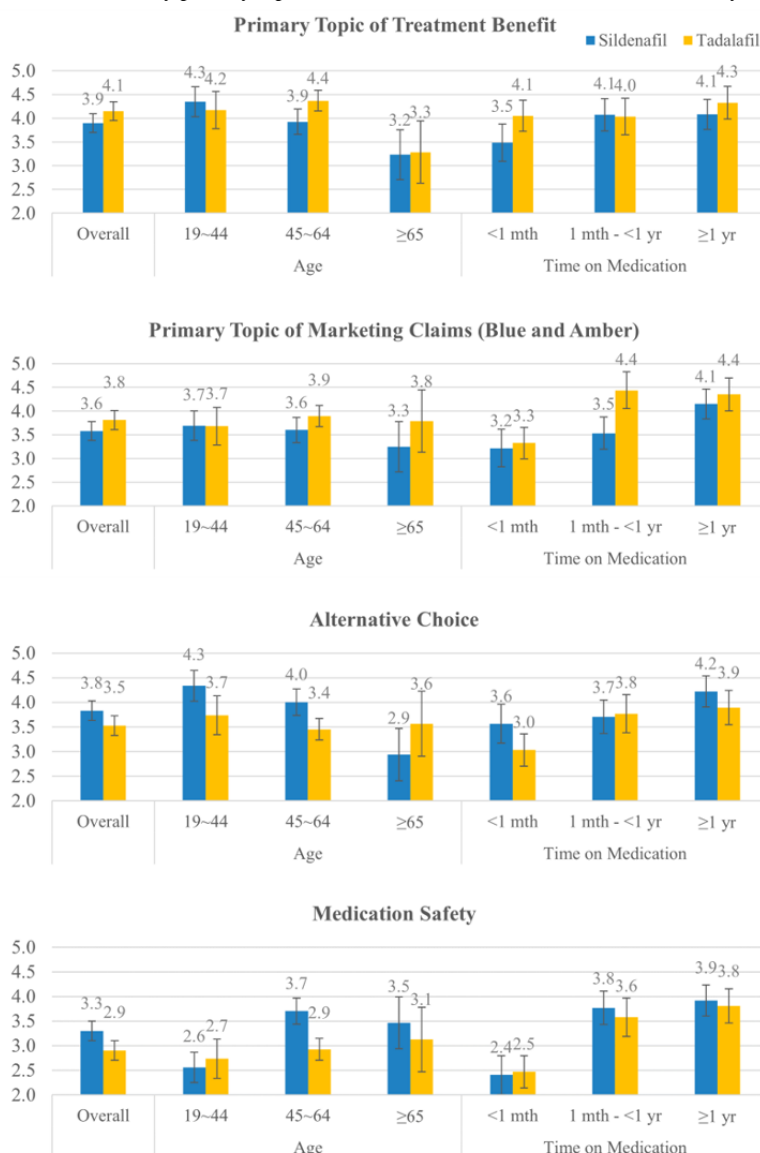
**Table 2.** Drug ratings of sildenafil by age and time on medication.

Variable	Overall rating, mean (SD)	Rating by age (years), mean (SD)			P value	Rating by time on medication, mean (SD)			
		19-44	45-64	≥65		<1 month	1 month to <1 year	≥1 year	P value
Sexual performance	3.94 (1.42)	4.32 (1.25)	4.02 (1.38)	3.44 (1.56)	.09	3.27 (1.73)	4.26 (1.33)	4.08 (1.16)	.02
Erection sustainability	3.85 (1.42)	4.37 (1.03)	3.82 (1.40)	2.86 (1.75)	.004	3.64 (1.42)	3.89 (1.49)	4.10 (1.37)	.50
Blue magic	3.58 (1.50)	3.70 (1.40)	3.60 (1.53)	3.25 (1.60)	.70	3.22 (1.53)	3.53 (1.55)	4.15 (1.29)	.04
Medication safety	3.30 (1.49)	2.56 (1.58)	3.70 (1.39)	3.46 (0.97)	.001	2.41 (1.42)	3.77 (1.45)	3.91 (1.09)	<.001
Alternative choices	3.83 (1.51)	4.33 (1.40)	4.00 (1.38)	2.94 (1.65)	.02	3.56 (1.75)	3.70 (1.64)	4.22 (1.09)	.28
Total	3.68 (1.48)	3.73 (1.53)	3.82 (1.42)	3.20 (1.52)	.004	3.16 (1.59)	3.88 (1.48)	4.08 (1.18)	<.001

**Table 3.** Drug ratings of tadalafil by age and time on medication.

Variable	Overall rating, mean (SD)	Rating by age (years), mean (SD)			P value	Rating by time on medication, mean (SD)			P value
		19-44	45-64	≥65		<1 month	1 month to <1 year	≥1 year	
Sexual performance	4.14 (1.29)	4.17 (1.15)	4.37 (1.12)	3.29 (1.70)	<.001	4.05 (1.27)	4.03 (1.47)	4.33 (1.13)	.48
Amber romance	3.81 (1.45)	3.68 (1.38)	3.89 (1.45)	3.79 (1.69)	.73	3.32 (1.53)	4.44 (1.12)	4.35 (0.99)	<.001
Mild medication safety	3.52 (1.52)	3.23 (1.61)	3.64 (1.49)	3.57 (1.43)	.30	3.07 (1.53)	4.06 (1.25)	4.00 (1.27)	<.001
Serious medication safety	2.44 (1.46)	2.34 (1.28)	2.47 (1.51)	2.50 (1.64)	.84	2.24 (1.39)	2.92 (1.42)	3.32 (1.73)	<.001
Alternative choices	3.53 (1.59)	3.74 (1.54)	3.45 (1.62)	3.56 (1.57)	.65	3.03 (1.82)	3.77 (1.41)	3.89 (1.37)	.005
Total	3.39 (1.59)	3.31 (1.54)	3.42 (1.61)	3.39 (1.62)	.65	2.91 (1.61)	3.87 (1.42)	4.01 (1.32)	<.001

**Figure 2.** Comparison of treatment satisfaction by primary topics between sildenafil and tadalafil. mth: month; yr: year.



## Discussion

### Principal Findings

NLP of patient medication reviews identified 5 topics per ED drug. Sildenafil had 2 topics on treatment benefit (*sexual*

*performance* and *erection sustainability*) and 1 topic on medication safety (*medication safety*). In contrast, tadalafil had 1 topic on treatment benefit (*sexual performance*) and 2 topics on medication safety (*mild medication safety* and *serious medication safety*).

*Erection sustainability* was additionally identified as a treatment benefit of sildenafil. Younger patients seemed to have received the most benefit from erection sustainability. The topic was more frequently mentioned (30/124, 24.2%) among younger patients than other age groups. Moreover, younger patients gave the highest satisfaction rating (4.37/5) when they mentioned *erection sustainability* as the primary topic compared with other primary topics. It has been known that sildenafil's marketing strategy was to increase its consumer base by appealing to younger adults [30,31]. To that end, the drug seller must have succeeded in incepting the concept that sildenafil enhances sexual performance, something younger adults desire, rather than treating the medical problem of ED prevalent among older adults [33-35].

It is a bit surprising that the reviews on tadalafil did not reveal *erection sustainability* as a topic considering that the drug remains longer in the blood compared with sildenafil. Evidently, erection sustainability must have meant how long an erection can last during sexual intercourse rather than how long the drug remains in the blood. Perhaps, tadalafil users were more concerned about erection readiness rather than erection sustainability. For this reason, older adults who desire erection readiness more than erection sustainability were more satisfied with tadalafil than with sildenafil [36,37].

The topic identification of patient medication reviews successfully uncovered the marketing claims of each drug, that is, the *amber romance* topic had a list of words like "last," "still," and "weekend," while the *blue magic* topic had a list of words like "hour," "dose," "minute," and "rapid." Eli Lilly, the tadalafil seller, knew that ED patients want sex to be more "natural" and therefore casted middle-aged actors in tadalafil commercials [38,39]. The commercials emphasized that the drug makes you ready whenever you feel like making love, which promotes romance over sexual acts. The seller even designed the pill to appear as a blown-up amber-colored heart. In contrast, Pfizer, the sildenafil seller, emphasized sexual performance over a romantic relationship. The seller incorporated a blue diamond shape into the pill design to make the drug look quite strong. These marketing claims are backed by some scientific evidence. The claim pertaining to *blue magic* is based on the pharmacokinetic property that the drug works rapidly and then clears out of the body with a half-life of 4 hours. On the other hand, the amber heart pill lasts for 3 days, which was promoted as a weekend pill where retaking the drug is not needed for successive sexual arousals for weekend romance.

The discovery that marketing claims are reflected in patient medication reviews suggests that ED drug users respond to marketing claims. The main goal of marketing is to identify who responds to commercial advertisements. In our study, the youngest age group was more satisfied with sildenafil than with tadalafil (score 3.73 vs 3.21). The youngest group was also more satisfied when *blue magic* was their primary topic rather than *amber romance*. These findings were reversed among the oldest group. Despite the greater uncertainty about the differentiation, both drugs seem to have successfully realized their respective marketing claims.

The numbers of topics related to medication safety were 2 for tadalafil and 1 for sildenafil. This indicates that safety concerning tadalafil has 2 subdimensions, one for *serious medication safety* and the other for *mild medication safety*, while sildenafil has 1 dimension called *medication safety*. Although tadalafil and sildenafil belong to the same class of PDE5 inhibitors, they clear out of the body differently; tadalafil lingers long in the blood, whereas sildenafil clears out of the body rapidly. Back pain and myalgia, which might be more prevalent among younger adults, result from PDE5 action [40]. Thus, it is likely that the lingering action of tadalafil could have aggravated the pain associated with PDE5 action [41].

Expectedly, patients who mentioned *serious medication safety* as the primary topic gave the lowest drug rating (2.44) compared with those who mentioned *medication safety* (3.30) or *mild medication safety* (3.52) as the primary topic. Among patients who had received ED drug therapy for less than 1 month, the primary topic of *serious medication safety* had the largest share (almost 40%). The share decreased to 10% among users who had used the drug for more than 1 year. It is worth noting that those who regarded *sexual performance* as the primary topic had a rating higher than 4.00 regardless of the time on tadalafil; however, among those with serious medication as the primary topic, the drug rating went up as the time on tadalafil increased. Logically, tadalafil users would stop taking the medication when they face a serious medication safety event. This explains why the proportion of patients who had used the ED drug for more than 1 year was lower for tadalafil than for sildenafil (187/919, 20.3% and 150/463, 32.4%, respectively).

Finally, the topic *alternative choices* was identified with regard to both drugs. It had a list of words like "Cialis," "trial," "another," and "generic" for sildenafil and words like "Viagra," "trial," "Levitra," and "generic" for tadalafil. It is certainly important for the patient to have access to alternative medications, especially since the high prices of ED drugs have been a burden on patients because of a lack of insurance coverage. In fact, patients frequently mentioned the generic versions that are 50 times cheaper than the branded pills [42]. Furthermore, the presence of the topic is aligned with the research in that one of the main reasons for risking to buy potentially counterfeit sexual stimulants, including Viagra and Cialis, is related to poor finance [43].

It is interesting why sildenafil users least frequently mentioned *alternative choices* as their primary topic, while tadalafil users mentioned it quite frequently (71/461, 15.4% vs 182/915, 19.9%). The alternative choices may not be as important to sildenafil users as they are to tadalafil users. Tadalafil users may have faced serious medication safety events (the largest share of primary topics: 244/915, 26.7%) and thus might have been motivated to talk about alternative choices. However, sildenafil users who less frequently (104/461, 22.6%) faced a medication safety event would have talked about it less frequently. Moreover, medication reviewers who had *alternative choices* as their primary topic were more satisfied with sildenafil than with tadalafil, except for the oldest group. The reviewers also gave better ratings to sildenafil than to tadalafil across multiple times on ED medication.

## Practice Implications

The identification of topics hidden in the patient reviews of ED drug therapy via topic modeling can have many applications. It can help evaluate whether the marketing claims have effectively targeted a specific group of people who desire a certain medication attribute for their health needs. It can also contribute to patient-centered care by informing health care providers of the different medication concerns facing individual patients taking ED drug therapy. Lastly, the study findings have documented the capabilities of topic modeling on SNS drug reviews in the areas of infodemiology/infoveillance of private and taboo topics. Topic modeling of ED drug reviews posted on SNSs can effectively reveal honest attitudes and opinions toward sexual needs not expressed in formal surveys. It could pave the way for topic modeling on SNS posts as an efficient social research tool to identify the needs of vulnerable populations whose opinions and orientations are not well accepted in society.

## Limitations

There may be biases that arise from using online reviews on social media. Online reviews may likely be posed by those who are eager to express their eccentricity. Therefore, it is likely they are not representative of the general public. In other words, the findings cannot be generalized to the public. However, the comparison between the 2 drugs may not have the limitation of selection bias since there appeared to be no systematic differences among the reviewers of the 2 ED drugs.

Patient medication experiences related to safety issues may have been exaggerated. It has been shown in previous research that a consumer's motivation to review a product is to inform others to avoid a negative experience [44,45]. Moreover, despite filtering the reviews, unidentified spam reviews might have gone undetected. Unfiltered spam reviews can affect the study results by intentionally giving false positive or malicious negative opinions about the drugs [46,47].

Naming each topic identified was done subjectively based on each list of words in each topic. Therefore, topic names may not capture all the minute nuances contained in each list. Furthermore, the researchers' subjectivity may have played an important role in extracting hidden topics since the number of topics is given by the authors. The optimal number of topics may vary based on specific criteria.

Despite the same data collection criteria, the number of patient medication reviews for sildenafil was almost half that for tadalafil. This may have resulted from the misaligned times between drug approval dates and SNS popularity [48]; drug reviews on SNSs were less popular when sildenafil was approved. In fact, ED was too sensitive to mention in public when sildenafil was first marketed. People became more comfortable with its discussion over time with continuous branding of ED as a medical problem to be treated [31,49]. Finally, tadalafil reviewers might have been more motivated to leave posts because they were more likely to mention medication safety than efficacy (ie, on medication safety, tadalafil had 2 topics while sildenafil had 1, and on efficacy, tadalafil had 1 topics while sildenafil had 2).

It is unlikely that the unbalanced number of patient medication reviews between the 2 drugs produced a bias in the study results. Because separate topic modeling was run for each drug set of reviews, the identification of topics would not be affected by the unbalanced number. However, it raises the question whether the number of sildenafil reviews was sufficient for topic modeling. It is reported that the sample size requirement for topic modeling varies with document characteristics, such as content heterogeneity and document length [50,51]. Patient medication reviews have a longer document length than typical tweets. They are also homogenous because they are from the patients who have taken medication for ED. It is reported that people with specific health problems provide informative and lengthy text data for health portals [52]. In addition, our study successfully identified 5 distinctive topics meeting the topic identification criteria [25]. Furthermore, a previous study successfully executed topic modeling based on less than 500 social reviews [53].

## Conclusion

The topics identified from patient medication reviews of ED drugs reflect drug labeling information, marketing claims, and comparative alternative choices facing patients in real-world practice. Topic modeling of natural language expressed in patient medication reviews can identify patient medication concerns, which are crucial for patient-centered prescription and medication counseling. Moreover, it supports that topic modeling on SNS posts is capable of uncovering hidden topics related to taboo or private behaviors.

---

## Acknowledgments

We thank the Research Institute of Pharmaceutical Science and College of Pharmacy, Seoul National University for supporting this study.

---

## Conflicts of Interest

None declared

---

## References

1. Pivovarov R, Perotte AJ, Grave E, Angiolillo J, Wiggins CH, Elhadad N. Learning probabilistic phenotypes from heterogeneous EHR data. *J Biomed Inform* 2015 Dec;58:156-165 [FREE Full text] [doi: [10.1016/j.jbi.2015.10.001](https://doi.org/10.1016/j.jbi.2015.10.001)] [Medline: [26464024](https://pubmed.ncbi.nlm.nih.gov/26464024/)]

2. Meng Y, Speier W, Ong M, Arnold CW. HCET: Hierarchical Clinical Embedding With Topic Modeling on Electronic Health Records for Predicting Future Depression. *IEEE J Biomed Health Inform* 2021 Apr;25(4):1265-1272. [doi: [10.1109/JBHI.2020.3004072](https://doi.org/10.1109/JBHI.2020.3004072)] [Medline: [32749975](https://pubmed.ncbi.nlm.nih.gov/32749975/)]
3. Spasic I, Button K. Patient Triage by Topic Modeling of Referral Letters: Feasibility Study. *JMIR Med Inform* 2020 Nov 06;8(11):e21252 [FREE Full text] [doi: [10.2196/21252](https://doi.org/10.2196/21252)] [Medline: [33155985](https://pubmed.ncbi.nlm.nih.gov/33155985/)]
4. Bisgin H, Liu Z, Fang H, Xu X, Tong W. Mining FDA drug labels using an unsupervised learning technique--topic modeling. *BMC Bioinformatics* 2011 Oct 18;12 Suppl 10:S11 [FREE Full text] [doi: [10.1186/1471-2105-12-S10-S11](https://doi.org/10.1186/1471-2105-12-S10-S11)] [Medline: [22166012](https://pubmed.ncbi.nlm.nih.gov/22166012/)]
5. Eysenbach G. Infodemiology and infoveillance: framework for an emerging set of public health informatics methods to analyze search, communication and publication behavior on the Internet. *J Med Internet Res* 2009 Mar 27;11(1):e11 [FREE Full text] [doi: [10.2196/jmir.1157](https://doi.org/10.2196/jmir.1157)] [Medline: [19329408](https://pubmed.ncbi.nlm.nih.gov/19329408/)]
6. Liu Q, Zheng Z, Zheng J, Chen Q, Liu G, Chen S, et al. Health Communication Through News Media During the Early Stage of the COVID-19 Outbreak in China: Digital Topic Modeling Approach. *J Med Internet Res* 2020 Apr 28;22(4):e19118 [FREE Full text] [doi: [10.2196/19118](https://doi.org/10.2196/19118)] [Medline: [32302966](https://pubmed.ncbi.nlm.nih.gov/32302966/)]
7. Boon-Itt S, Skunkan Y. Public Perception of the COVID-19 Pandemic on Twitter: Sentiment Analysis and Topic Modeling Study. *JMIR Public Health Surveill* 2020 Nov 11;6(4):e21978 [FREE Full text] [doi: [10.2196/21978](https://doi.org/10.2196/21978)] [Medline: [33108310](https://pubmed.ncbi.nlm.nih.gov/33108310/)]
8. Doogan C, Buntine W, Linger H, Brunt S. Public Perceptions and Attitudes Toward COVID-19 Nonpharmaceutical Interventions Across Six Countries: A Topic Modeling Analysis of Twitter Data. *J Med Internet Res* 2020 Sep 03;22(9):e21419 [FREE Full text] [doi: [10.2196/21419](https://doi.org/10.2196/21419)] [Medline: [32784190](https://pubmed.ncbi.nlm.nih.gov/32784190/)]
9. Schück S, Foulquié P, Mebarki A, Faviez C, Khadhar M, Texier N, et al. Concerns Discussed on Chinese and French Social Media During the COVID-19 Lockdown: Comparative Infodemiology Study Based on Topic Modeling. *JMIR Form Res* 2021 Apr 05;5(4):e23593 [FREE Full text] [doi: [10.2196/23593](https://doi.org/10.2196/23593)] [Medline: [33750736](https://pubmed.ncbi.nlm.nih.gov/33750736/)]
10. Hu T, Wang S, Luo W, Zhang M, Huang X, Yan Y, et al. Revealing Public Opinion Towards COVID-19 Vaccines With Twitter Data in the United States: Spatiotemporal Perspective. *J Med Internet Res* 2021 Sep 10;23(9):e30854 [FREE Full text] [doi: [10.2196/30854](https://doi.org/10.2196/30854)] [Medline: [34346888](https://pubmed.ncbi.nlm.nih.gov/34346888/)]
11. Luo C, Ji K, Tang Y, Du Z. Exploring the Expression Differences Between Professionals and Laypeople Toward the COVID-19 Vaccine: Text Mining Approach. *J Med Internet Res* 2021 Aug 27;23(8):e30715 [FREE Full text] [doi: [10.2196/30715](https://doi.org/10.2196/30715)] [Medline: [34346885](https://pubmed.ncbi.nlm.nih.gov/34346885/)]
12. Bae J, Han N, Song M. Twitter Issue Tracking System by Topic Modeling Techniques. *Journal of Intelligence and Information Systems* 2014 Jun 30;20(2):109-122 [FREE Full text] [doi: [10.13088/jiis.2014.20.2.109](https://doi.org/10.13088/jiis.2014.20.2.109)]
13. Sharma E, Saha K, Ernala S, Ghoshal S, De Choudhury M. Analyzing Ideological Discourse on Social Media: A Case Study of the Abortion Debate. In: *CSS 2017: Proceedings of the 2017 International Conference of The Computational Social Science Society of the Americas*. 2017 Presented at: International Conference of The Computational Social Science Society of the Americas; October 19-22, 2017; Santa Fe, NM p. 1-8. [doi: [10.1145/3145574.3145577](https://doi.org/10.1145/3145574.3145577)]
14. Xue J, Chen J, Gelles R. Using Data Mining Techniques to Examine Domestic Violence Topics on Twitter. *Violence and Gender* 2019 Jun;6(2):105-114 [FREE Full text] [doi: [10.1089/vio.2017.0066](https://doi.org/10.1089/vio.2017.0066)]
15. Kang J, Kim S, Roh S. A Topic Modeling Analysis for Online News Article Comments on Nurses' Workplace Bullying. *J Korean Acad Nurs* 2019 Dec;49(6):736-747 [FREE Full text] [doi: [10.4040/jkan.2019.49.6.736](https://doi.org/10.4040/jkan.2019.49.6.736)] [Medline: [31932568](https://pubmed.ncbi.nlm.nih.gov/31932568/)]
16. Ong AD, Weiss DJ. The Impact of Anonymity on Responses to Sensitive Questions. *J Appl Social Psychol* 2000 Aug;30(8):1691-1708. [doi: [10.1111/j.1559-1816.2000.tb02462.x](https://doi.org/10.1111/j.1559-1816.2000.tb02462.x)]
17. Capurro D, Cole K, Echavarría MI, Joe J, Neogi T, Turner AM. The Use of Social Networking Sites for Public Health Practice and Research: A Systematic Review. *J Med Internet Res* 2014 Mar 14;16(3):e79 [FREE Full text] [doi: [10.2196/jmir.2679](https://doi.org/10.2196/jmir.2679)] [Medline: [24642014](https://pubmed.ncbi.nlm.nih.gov/24642014/)]
18. Wei S, Ma M, Wen X, Wu C, Zhu G, Zhou X. Online Public Attention Toward Premature Ejaculation in Mainland China: Infodemiology Study Using the Baidu Index. *J Med Internet Res* 2021 Aug 26;23(8):e30271 [FREE Full text] [doi: [10.2196/30271](https://doi.org/10.2196/30271)] [Medline: [34435970](https://pubmed.ncbi.nlm.nih.gov/34435970/)]
19. Park SH, Hong SH. Identification of Primary Medication Concerns Regarding Thyroid Hormone Replacement Therapy From Online Patient Medication Reviews: Text Mining of Social Network Data. *J Med Internet Res* 2018 Oct 24;20(10):e11085 [FREE Full text] [doi: [10.2196/11085](https://doi.org/10.2196/11085)] [Medline: [30355555](https://pubmed.ncbi.nlm.nih.gov/30355555/)]
20. Han N, Oh JM, Kim I. Assessment of adverse events related to anti-influenza neuraminidase inhibitors using the FDA adverse event reporting system and online patient reviews. *Sci Rep* 2020 Feb 20;10(1):3116 [FREE Full text] [doi: [10.1038/s41598-020-60068-5](https://doi.org/10.1038/s41598-020-60068-5)] [Medline: [32080337](https://pubmed.ncbi.nlm.nih.gov/32080337/)]
21. Benton A, Ungar L, Hill S, Hennessy S, Mao J, Chung A, et al. Identifying potential adverse effects using the web: a new approach to medical hypothesis generation. *J Biomed Inform* 2011 Dec;44(6):989-996 [FREE Full text] [doi: [10.1016/j.jbi.2011.07.005](https://doi.org/10.1016/j.jbi.2011.07.005)] [Medline: [21820083](https://pubmed.ncbi.nlm.nih.gov/21820083/)]
22. Walsh J, Cave J, Griffiths F. Spontaneously Generated Online Patient Experience of Modafinil: A Qualitative and NLP Analysis. *Front Digit Health* 2021 Feb 17;3:598431 [FREE Full text] [doi: [10.3389/fgdth.2021.598431](https://doi.org/10.3389/fgdth.2021.598431)] [Medline: [34713085](https://pubmed.ncbi.nlm.nih.gov/34713085/)]
23. WebMD. URL: <https://www.webmd.com/> [accessed 2022-02-16]
24. Ask a Patient. URL: <https://www.askapatient.com/> [accessed 2022-02-16]

25. Blei DM, Ng AY, Jordan MI. Latent Dirichlet Allocation. *Journal of Machine Learning Research* 2003;3:993-1022 [[FREE Full text](#)]
26. Cao J, Xia T, Li J, Zhang Y, Tang S. A density-based method for adaptive LDA model selection. *Neurocomputing* 2009 Mar;72(7-9):1775-1781 [[FREE Full text](#)] [doi: [10.1016/j.neucom.2008.06.011](https://doi.org/10.1016/j.neucom.2008.06.011)]
27. Sievert C, Shirley K. LDavis: A method for visualizing and interpreting topics. In: *Proceedings of the Workshop on Interactive Language Learning, Visualization, and Interfaces*. 2014 Presented at: Workshop on Interactive Language Learning, Visualization, and Interfaces; June 2014; Baltimore, MD p. 63-70. [doi: [10.3115/v1/W14-3110](https://doi.org/10.3115/v1/W14-3110)]
28. Maier D, Waldherr A, Miltner P, Wiedemann G, Niekler A, Keinert A, et al. Applying LDA Topic Modeling in Communication Research: Toward a Valid and Reliable Methodology. *Communication Methods and Measures* 2018 Feb 16;12(2-3):93-118. [doi: [10.1080/19312458.2018.1430754](https://doi.org/10.1080/19312458.2018.1430754)]
29. Drugs@FDA: FDA-Approved Drugs. U.S. Food and Drug Administration. URL: <https://www.accessdata.fda.gov/scripts/cder/daf/> [accessed 2022-02-16]
30. Lexchin J. Bigger and better: how Pfizer redefined erectile dysfunction. *PLoS Med* 2006 Apr 11;3(4):e132 [[FREE Full text](#)] [doi: [10.1371/journal.pmed.0030132](https://doi.org/10.1371/journal.pmed.0030132)] [Medline: [16597172](https://pubmed.ncbi.nlm.nih.gov/16597172/)]
31. Gesser-Edelsburg A, Hijazi R. The Magic Pill: The Branding of Impotence and the Positioning of Viagra as Its Solution through Edutainment. *J Health Commun* 2018 Feb 13;23(3):281-290. [doi: [10.1080/10810730.2018.1436623](https://doi.org/10.1080/10810730.2018.1436623)] [Medline: [29436964](https://pubmed.ncbi.nlm.nih.gov/29436964/)]
32. Wienke C. Male sexuality, medicalization, and the marketing of Cialis and Levitra. *Sex Cult* 2005 Dec;9(4):29-57. [doi: [10.1007/s12119-005-1001-1](https://doi.org/10.1007/s12119-005-1001-1)]
33. Moynihan R. The Rise of Viagra: How the Little Blue Pill Changed Sex in America. *BMJ* 2005 Feb 17;330(7488):424. [doi: [10.1136/bmj.330.7488.424](https://doi.org/10.1136/bmj.330.7488.424)]
34. Younger men lead surge in Viagra use, study reveals. *EurekaAlert*. 2004. URL: [https://www.eurekaalert.org/pub\\_releases/2004-08/es-yml080204.php](https://www.eurekaalert.org/pub_releases/2004-08/es-yml080204.php) [accessed 2022-08-11]
35. Jagged Little (Blue) Pill. *The Atlantic*. 2018. URL: <https://www.theatlantic.com/entertainment/archive/2018/03/20-years-of-viagra/556343/> [accessed 2022-08-11]
36. Hanson-Divers C, Jackson SE, Lue TF, Crawford SY, Rosen RC. Health outcomes variables important to patients in the treatment of erectile dysfunction. *J Urol* 1998 May;159(5):1541-1547. [doi: [10.1097/00005392-199805000-00037](https://doi.org/10.1097/00005392-199805000-00037)] [Medline: [9554350](https://pubmed.ncbi.nlm.nih.gov/9554350/)]
37. Doggrell S. Do vardenafil and tadalafil have advantages over sildenafil in the treatment of erectile dysfunction? *Int J Impot Res* 2007 Dec 21;19(3):281-295. [doi: [10.1038/sj.ijir.3901525](https://doi.org/10.1038/sj.ijir.3901525)] [Medline: [17183346](https://pubmed.ncbi.nlm.nih.gov/17183346/)]
38. Viagra's success brings challenge from two rivals. *Pittsburgh Post-Gazette*. 2003. URL: <https://www.post-gazette.com/business/businessnews/2003/05/02/Viagra-s-success-brings-challenge-from-two-rivals/stories/200305020014> [accessed 2022-02-16]
39. Conaglen HM, Conaglen JV. Investigating women's preference for sildenafil or tadalafil use by their partners with erectile dysfunction: the partners' preference study. *J Sex Med* 2008 May;5(5):1198-1207. [doi: [10.1111/j.1743-6109.2008.00774.x](https://doi.org/10.1111/j.1743-6109.2008.00774.x)] [Medline: [18312284](https://pubmed.ncbi.nlm.nih.gov/18312284/)]
40. Seftel AD, Farber J, Fletcher J, Deeley MC, Elion-Mboussa A, Hoover A, et al. A three-part study to investigate the incidence and potential etiologies of tadalafil-associated back pain or myalgia. *Int J Impot Res* 2005;17(5):455-461. [doi: [10.1038/sj.ijir.3901374](https://doi.org/10.1038/sj.ijir.3901374)] [Medline: [16034469](https://pubmed.ncbi.nlm.nih.gov/16034469/)]
41. Taylor J, Baldo OB, Storey A, Cartledge J, Eardley I. Differences in side-effect duration and related bother levels between phosphodiesterase type 5 inhibitors. *BJU Int* 2009 May;103(10):1392-1395. [doi: [10.1111/j.1464-410X.2008.08328.x](https://doi.org/10.1111/j.1464-410X.2008.08328.x)] [Medline: [19154494](https://pubmed.ncbi.nlm.nih.gov/19154494/)]
42. Skinner G. How to Get Generic Viagra. *Consumer Reports*. 2017. URL: <https://www.consumerreports.org/prescription-drugs/how-to-get-generic-viagra/> [accessed 2022-02-15]
43. Assi S, Thomas J, Haffar M, Osselton D. Exploring Consumer and Patient Knowledge, Behavior, and Attitude Toward Medicinal and Lifestyle Products Purchased From the Internet: A Web-Based Survey. *JMIR Public Health Surveill* 2016 Jul 18;2(2):e34 [[FREE Full text](#)] [doi: [10.2196/publichealth.5390](https://doi.org/10.2196/publichealth.5390)] [Medline: [27430264](https://pubmed.ncbi.nlm.nih.gov/27430264/)]
44. Dellarcas C, Narayan R, Smith R. What motivates consumers to review a product online? A study of the product-specific antecedents of online movie reviews. *Citeseerx*. URL: <https://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.452.6139&rep=rep1&type=pdf> [accessed 2022-02-16]
45. Hennig-Thurau T, Gwinner K, Walsh G, Gremler D. Electronic word-of-mouth via consumer-opinion platforms: What motivates consumers to articulate themselves on the Internet? *Journal of Interactive Marketing* 2004 Jan;18(1):38-52 [[FREE Full text](#)] [doi: [10.1002/dir.10073](https://doi.org/10.1002/dir.10073)]
46. Jindal N, Liu B. Review spam detection. In: *WWW '07: Proceedings of the 16th International Conference on World Wide Web*. 2007 Presented at: 16th International Conference on World Wide Web; May 8-12, 2007; Banff, Alberta, Canada. [doi: [10.1145/1242572.1242759](https://doi.org/10.1145/1242572.1242759)]
47. Crawford M, Khoshgoftaar TM, Prusa JD, Richter AN, Al Najada H. Survey of review spam detection using machine learning techniques. *Journal of Big Data* 2015 Oct 5;2(1):23. [doi: [10.1186/s40537-015-0029-9](https://doi.org/10.1186/s40537-015-0029-9)]



48. Sprague D. The History Of Online Reviews And How They Have Evolved. Shopper Approved. 2019. URL: <https://5stars.shopperapproved.com/the-history-and-evolution-of-online-reviews/> [accessed 2022-02-16]
49. Parsons P. Integrating ethics with strategy: analyzing disease - branding. Corporate Communications: An International Journal 2007;12(3):267-279. [doi: [10.1108/13563280710776860](https://doi.org/10.1108/13563280710776860)]
50. Naushan H. Topic Modeling with Latent Dirichlet Allocation. Towards Data Science. 2020. URL: <https://towardsdatascience.com/topic-modeling-with-latent-dirichlet-allocation-e7ff75290f8> [accessed 2022-02-16]
51. Maier D, Niekler A, Wiedemann G, Stoltenberg D. How Document Sampling and Vocabulary Pruning Affect the Results of Topic Models. Computational Communication Research 2020;2(2):139-152. [doi: [10.5117/CCR2020.2.001.MAIE](https://doi.org/10.5117/CCR2020.2.001.MAIE)]
52. Eysenbach G. Credibility of Health Information and Digital Media: New Perspectives and Implications for Youth. MacArthur Foundation Digital Media and Learning Initiative 2008:123-154. [doi: [10.1162/dmal.9780262562324.123](https://doi.org/10.1162/dmal.9780262562324.123)]
53. Xianghua F, Guo L, Yanyan G, Zhiqiang W. Multi-aspect sentiment analysis for Chinese online social reviews based on topic modeling and HowNet lexicon. Knowledge-Based Systems 2013 Jan;37:186-195. [doi: [10.1016/j.knosys.2012.08.003](https://doi.org/10.1016/j.knosys.2012.08.003)]

## Abbreviations

**ED:** erectile dysfunction  
**FDA:** Food and Drug Administration  
**LDA:** latent Dirichlet allocation  
**NLP:** natural language processing  
**PDE5:** phosphodiesterase type 5  
**PRO:** patient-reported outcome  
**SNS:** social network service

*Edited by G Eysenbach; submitted 06.08.21; peer-reviewed by A Mavragani; comments to author 27.08.21; revised version received 22.10.21; accepted 17.11.21; published 28.02.22.*

*Please cite as:*

*Kim M, Noh Y, Yamada A, Hong SH*

*Comparison of the Erectile Dysfunction Drugs Sildenafil and Tadalafil Using Patient Medication Reviews: Topic Modeling Study*  
*JMIR Med Inform 2022;10(2):e32689*

*URL: <https://medinform.jmir.org/2022/2/e32689>*

*doi: [10.2196/32689](https://doi.org/10.2196/32689)*

*PMID: [35225813](https://pubmed.ncbi.nlm.nih.gov/35225813/)*

©Maryanne Kim, Youran Noh, Akihiko Yamada, Song Hee Hong. Originally published in JMIR Medical Informatics (<https://medinform.jmir.org>), 28.02.2022. This is an open-access article distributed under the terms of the Creative Commons Attribution License (<https://creativecommons.org/licenses/by/4.0/>), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Medical Informatics, is properly cited. The complete bibliographic information, a link to the original publication on <https://medinform.jmir.org/>, as well as this copyright and license information must be included.

Original Paper

# Electricity, Computing Hardware, and Internet Infrastructures in Health Facilities in Sierra Leone: Field Mapping Study

Emeka Chukwu<sup>1</sup>, BEng, MSc; Lalit Garg<sup>1</sup>, BE, MSc, PhD; Edward Foday<sup>2</sup>, BA, MSc, MPhil; Abdul Konomanyi<sup>3</sup>, BEng, MSc; Royston Wright<sup>4</sup>, BSc, MA, MSc; Francis Smart<sup>2</sup>, MD, MPH

<sup>1</sup>Department of Computer Information Systems, Faculty of Information Communications Technology, University of Malta, Msida, Malta

<sup>2</sup>Directorate of Policy, Planning, and Information, Ministry of Health and Sanitation, Freetown, Sierra Leone

<sup>3</sup>Directorate of eGovernment, Ministry of Information and Communication, Freetown, Sierra Leone

<sup>4</sup>Monitoring and Evaluation Unit, Health and Nutrition, United Nations Children's Fund, Freetown, Sierra Leone

**Corresponding Author:**

Emeka Chukwu, BEng, MSc

Department of Computer Information Systems

Faculty of Information Communications Technology

University of Malta

Tal-Qroqq

Msida, 2080

Malta

Phone: 356 99330888

Email: [nnaemeka\\_ec@hotmail.com](mailto:nnaemeka_ec@hotmail.com)

## Abstract

**Background:** Years of health information system investment in many countries have facilitated service delivery, surveillance, reporting, and monitoring. Electricity, computing hardware, and internet networks are vital for health facility-based information systems. Availability of these infrastructures at health facilities is crucial for achieving national digital health visions.

**Objective:** The aim of this study was to gain insight into the state of computing hardware, electricity, and connectivity infrastructure at health facilities in Sierra Leone using a representative sample.

**Methods:** Stratified sampling of 72 (out of 1284) health facilities distributed in all districts of Sierra Leone was performed, factoring in the rural-urban divide, digital health activity, health facility type, and health facility ownership. Enumerators visited each health facility over a 2-week period.

**Results:** Among the 72 surveyed health facilities, 59 (82%) do not have institutionally provided internet. Among the 15 Maternal and Child Health Posts, as a type of primary health care unit (PHU), 9 (60%) use solar energy as their only electricity source and the other 6 (40%) have no electricity source. Similarly, among the 13 hospitals, 5 (38%) use a generator as a primary electricity source. All hospitals have at least one functional computer, although only 7 of the 13 hospitals have four or more functional computers. Similarly, only 2 of the 59 (3%) PHUs have one computer each, and 37 (63%) of the PHUs have one tablet device each. We consider this health care computing infrastructure mapping to be representative with a 95% confidence level within an 11% margin of error. Two-thirds of the PHUs have only alternate solar electricity, only 10 of the 72 surveyed health facilities have functional official internet, and most use suboptimal computing hardware. Overall, 43% of the surveyed health facilities believe that inadequate electricity is the biggest threat to digitization. Similarly, 16 (22%) of the 72 respondents stated that device theft is a primary hindrance to digitization.

**Conclusions:** Electricity provision for off-electricity-grid health facilities using alternative and renewable energy sources is emerging. The current trend where GSM (Global System for Mobile Communication) service providers provide the internet to all health facilities may change to other promising alternatives. This study provides evidence of the critical infrastructure gaps in health facilities in Sierra Leone.

(*JMIR Med Inform* 2022;10(2):e30040) doi:[10.2196/30040](https://doi.org/10.2196/30040)

**KEYWORDS**

digital health; mHealth; eHealth infrastructure; health ICT; SpaceX; connectivity; Sierra Leone; rural-urban divide; rural areas; internet

## Introduction

### Background

Globally, health systems technology infrastructure has been a topic of constant debate. Technology infrastructure can help fast-track attainment of the global Sustainable Development Goal targets. In 2004, the European eHealth Action Plan considered technology infrastructure as critical to deployed solutions [1]. Liu et al [2] explored the challenges and solutions of deploying eHealth infrastructure. Electricity was listed as one of the essential amenities by the World Health Organization (WHO) handbook on monitoring building blocks of the health system [3]. Omotosho et al [4] surveyed the current state of information and communications technology (ICT) and related infrastructure supporting eHealth deployment in Africa. However, their desk survey covered computing hardware but not electricity. Sierra Leone has approximately 100 megawatts of electricity installed, with electricity per capita estimated at 34 kilowatt hours [5,6]. As of 2019, an estimated 22.7% of Sierra Leoneans have access to electricity, with only 1.5% of those in rural areas having electricity access [5,6]. There are three leading mobile service providers in Sierra Leone and broadband internet utilization remains low [7].

Sierra Leone has 1284 health facilities, including 24 district hospitals spread across the 13 health care districts [8]. The health systems are split between Primary Healthcare Units (PHUs) and hospitals. The PHUs are classified into Community Health Centers (CHCs), Community Health Posts (CHPs), and Maternal and Child Health Posts (MCHPs) [9]. The MCHPs operate at the village level, which serve less than 5000 people and are mainly staffed by Maternal and Child Health (MCH) aides who provide mainly maternal health services [10]. The CHPs operate at the town level and are situated to serve 5000 to 10,000 people. The CHPs are staffed by a community health nurse and MCH aides. In addition to services provided by MCH, they also provide disease prevention and control. The CHCs cover the chiefdoms, with an estimated population of 10,000 to 20,000 people. The CHCs are staffed by Community Health Officers and those found in CHPs. The CHCs conduct disease surveillance services. Hospitals are located to meet service needs, with at least one per district. Health facilities can be either public or privately owned, although the majority are public. Investments in health information systems (HISs) has resulted in regular health facility service delivery and disease surveillance reporting, using the District Health Information System in recent years [8,11]. Other HIS data sources include health surveys, birth registrations, census, and health resource tracking (eg, health accounts) [3].

Approximately 98% of health facilities consistently submit aggregated service delivery data to the central repository [9].

In 2017, the Directorate of Planning, Policy, and Information (DPPI) at the Ministry of Health and Sanitation (MoHS) inaugurated the eHealth coordination hub to govern the systematic application of digital health solutions for health systems improvement through data [12]. This culminated in the launch of the first national digital health strategy 2018-2023 [9]. The vision of the national digital health strategy is to guarantee universal health coverage using ICT. According to this strategy, service delivery data and disease surveillance data are collected and aggregated using a mixture of paper and digital health tools. This health information flows from the community, PHUs, or hospitals up to the district. Universal health coverage is one strategy that can ensure meeting global health care targets for different health domains.

### Study Objective

The eHealth coordination hub commissioned a mapping of the digital health-enabling environment components in Sierra Leone's health facilities in January 2019. This study unearths the state of digital health infrastructure as defined by the WHO-International Telecommunication Union eHealth strategy development toolkit [13] using a representative sample. The infrastructure (or information structures) that support collection, processing, and knowledge-mining of individualized patient data can be categorized as connectivity, computing hardware, and electricity. This study does not discuss other architectural (nonphysical) information structures such as standards and interoperability components. Instead, this study builds upon previous success to provide evidence of the linkage between the availability of these infrastructures and the availability of individualized digital health data in support of the national vision.

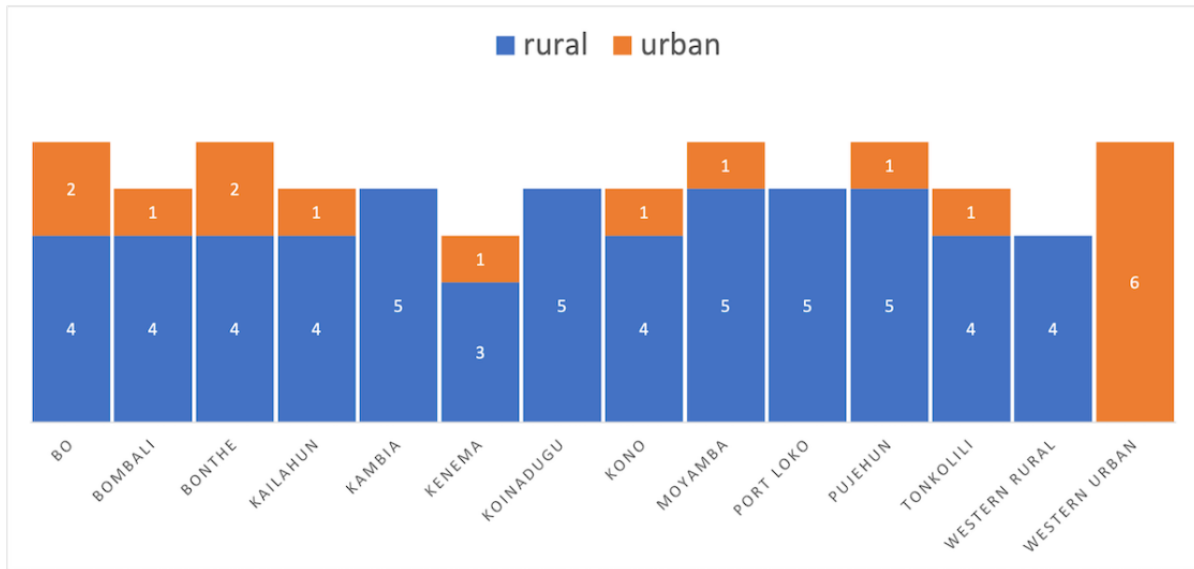
## Methods

### Health Facilities Sampling Strategy

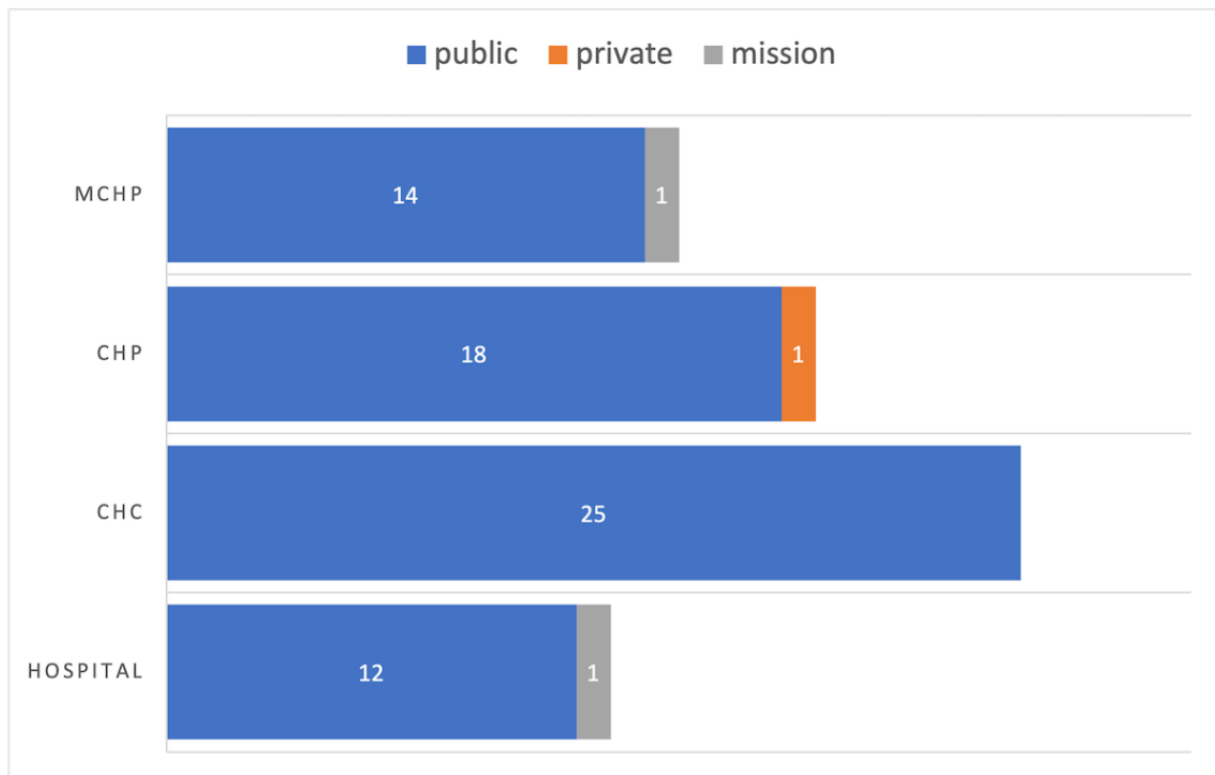
We used a stratified sampling strategy where each of the 13 health care districts in Sierra Leone was purposefully targeted for 5 or more health facilities. In total, 72 of the 1284 health facilities in Sierra Leone were selected for this mapping exercise, including 17 urban and 55 rural health facilities (Figure 1).

The margin of error for the 72 samples was calculated to be 11% to yield 95% confidence that the sample is representative [14]. The health facilities surveyed and their distribution, by ownership and type, are given in Figure 2. Among the 72 health facilities sampled, 69 (96%) were in the public sector, which aligns with the national digital health strategy and the country's current state of health facility distribution.

**Figure 1.** Distribution of surveyed health facilities by all districts and economic characteristics.



**Figure 2.** Health facilities surveyed, by ownership and type. MCHP: Maternal and Child Health Post; CHP: Community Health Post; CHC: Community Health Center.



Health facilities were initially classified as either urban or rural for spread and inclusion, based on information from the DPPI at MoHS, working in conjunction with the respective District Health Management Team heads. Health facilities were further classified according to the level of their digital health activity. For this mapping exercise only, health facilities were classified into three groups according to low, medium, or high digital health activity based on having no digital health solution, one or two digital health solutions, and three or more digital health solutions, respectively. We sampled a minimum of 5 health facilities per district, selecting 2 each from urban and rural location in each district, as precategorized. Each district

prioritized at least one facility with high digital health activity, followed by at least one health facility with medium activity, and finally one with no activity. Because each district had a minimum of one district hospital, only one district hospital was selected in each district irrespective of their digital health activity. Additional health facilities were selected by repeating this selection technique until the desired number was reached in each district. In a situation for which some categories did not exist (eg, no high-activity digital health facility), the required numbers were filled in with other categories.

### Data Collection and Analysis

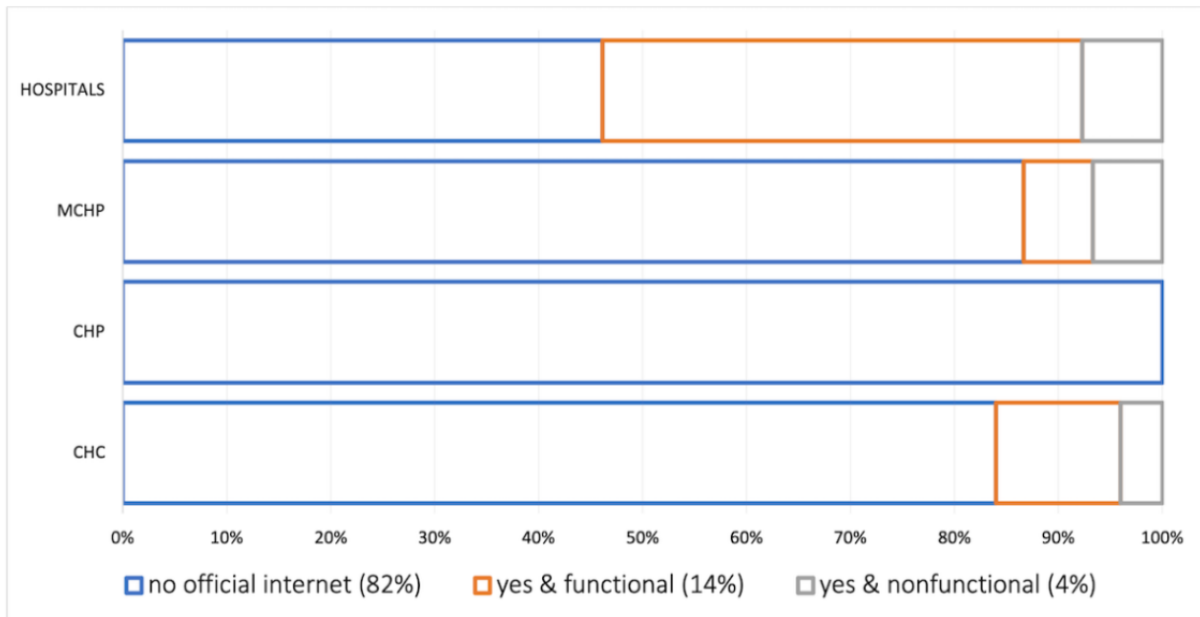
Ten study enumerators were recruited, trained, and deployed for this exercise in January 2019. The study enumerators visited assigned health facilities and interviewed the head of the health facilities while observing for infrastructures. The enumerators collected data using mobile forms, which were exported into an Excel spreadsheet. The Excel data were then analyzed with the *pandas* and *matplotlib* Python libraries [15].

### Results

#### Internet Connectivity

Respondents at 19 of the 72 health facilities surveyed reported having unofficial, private internet access at work, including at 6 of the 13 district hospitals surveyed. Approximately 90% (53/59) of the primary health care facilities surveyed did not have official institutional-provided internet. Likewise, half of the hospitals did not have official internet, as illustrated in Figure 3.

**Figure 3.** Percentage of health facilities that have institutionally provided internet. MCHP: Maternal and Child Health Post; CHP: Community Health Post; CHC: Community Health Center.



#### Electricity

All hospitals surveyed had an electric power source; 6 of them had a national utility grid, 5 had generators, and 1 had solar panels as the primary source. The detailed distribution of the number and type of primary electricity sources according to the different types of health facilities is provided in Table 1.

Approximately half of the PHUs surveyed did not have an alternative electricity supply source. All hospitals had one or more alternative electricity supply sources. Approximately half of the PHUs use their primary electricity source (ie, solar) for one purpose only, and the other half use the electricity for all health facility needs.

**Table 1.** Primary electricity sources of the surveyed health facilities (N=72).

Facility type	National utility	Generator	Solar	No electricity	Did not specify
Hospital	6	5	1	0	1
CHC <sup>a</sup>	6	1	17	1	0
CHP <sup>b</sup>	4	0	13	2	0
MCHP <sup>c</sup>	0	0	9	6	0

<sup>a</sup>CHC: Community Health Center.

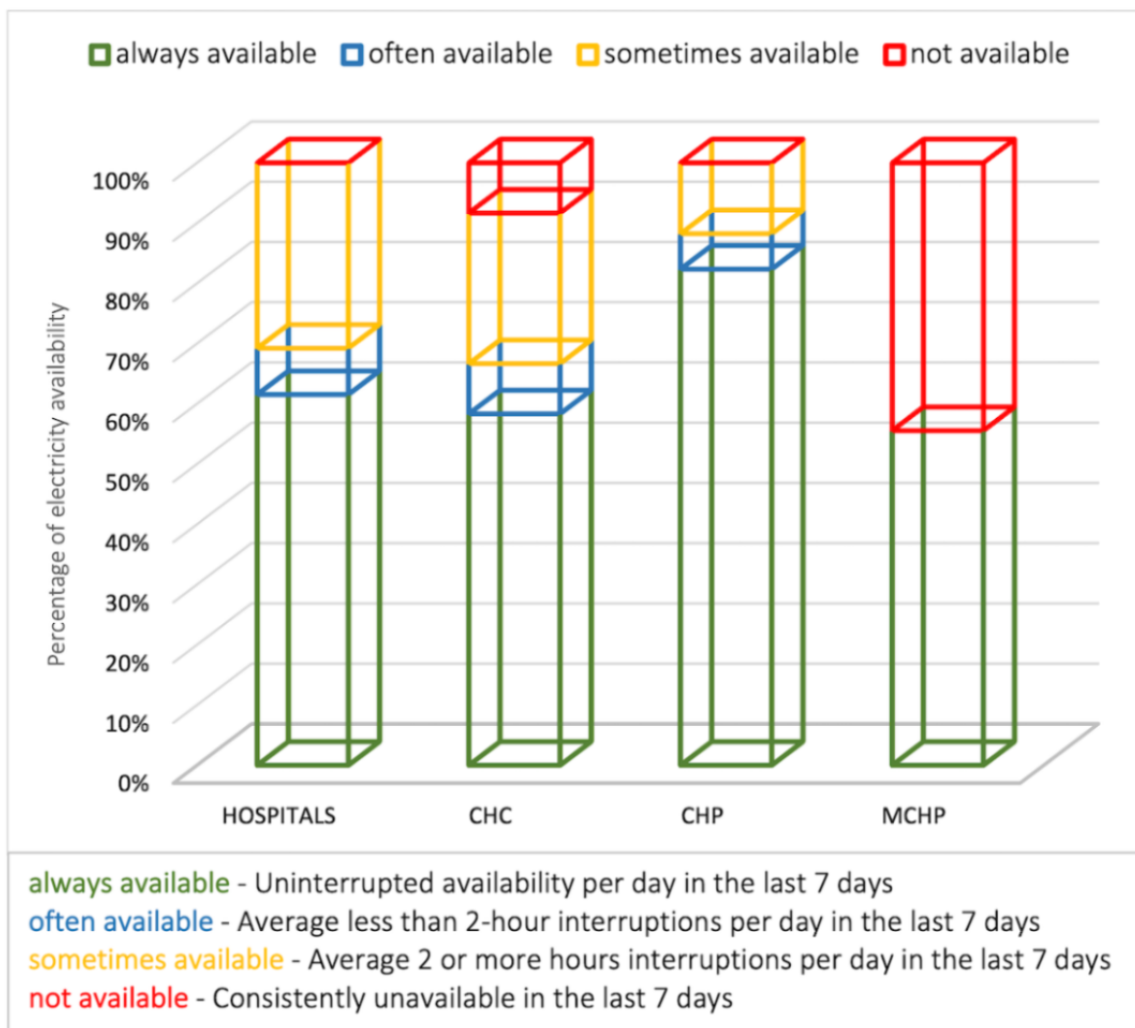
<sup>b</sup>CHP: Community Health Post.

<sup>c</sup>MCHP: Maternal and Child Health Post.

Respondents at health facilities with primary electricity sources were asked how long electricity was available at their health facilities, using their recall about availability in the 7 days before the survey. Approximately two-thirds of the hospitals surveyed indicated that they had an uninterrupted power supply in the

previous 7 days. Moreover, none of the hospitals surveyed reported consistently unavailable electricity. The results for the PHUs were mixed, with the MCHPs having the worst findings (see Figure 4).

**Figure 4.** Duration of electricity availability at health facilities. CHC: Community Health Center; CHP: Community Health Post; MCHP: Maternal and Child Health Post.



Eleven of the 13 hospitals had a generator, and most of the generators at these hospitals were functional. One hospital had a nonfunctional generator and one had no generator. At the time of the survey visit, only 8 of the 11 hospitals with a functional generator had fuel in the event of a power outage. Similarly, only 1 of the 6 PHUs with functional generators had fuel at the survey period. Nine of the PHUs with solar as the primary electricity source had a partially functional solar and inverter system.

### Computing Hardware

All 13 hospitals had at least one functional computer at the time of enumerator visits. Moreover, only 2 PHUs had a functional computer. However, only 7 hospitals had 4 or more functional computers. One hospital had 15 functional computers. Thirty-seven of the health facilities with tablet-based digital health solutions had only one tablet. Other statistics from the survey showed that only 8 PHUs had one smartphone and one hospital had eight smartphones. One PHU had two feature phones and one hospital had one feature phone. Six PHUs had one basic phone (dumbphone) that could only be used for calls and SMS text messaging. Three hospitals had internet modems. One hospital had one modem, one hospital had two modems, and one hospital had three modems.

## Discussion

### Principal Findings

Infrastructure in support of information systems is at the core of the success of an HIS [16]. Individualized care can be better optimized for data use when seamless HIS electricity, computing devices, and internet-network infrastructures are available. This study mapped 72 health facilities, which were first divided into subgroups by health facility type, ownership, and rural-urban distribution. The selection was further stratified to consider digital services and applications use in the health facility for adequate representation. This sample gave a 95% confidence level with an 11% margin of error. This means that the findings in this report are statistically generalizable. This study focused on Sierra Leone, which is considered an excellent example of a low- and middle-income country (LMIC), although we are aware that other LMICs may vary slightly in their characteristics [17].

The trends from our findings show the increasing use of solar solutions for PHUs, which are located mostly in rural locations. These rural facilities are often disconnected from the national electricity utility grid. This is a crucial lesson Sierra Leone shares with other LMICs with similar infrastructure deficits

[18]. Some MCHPs still do not have any electricity source, and any health facility digitization depends on electricity. Renewable energy sources are bridging these gaps. Our findings show that 66% (39/59) of PHUs use solar as their primary electric energy source and 15% (9/59) have had no electricity source. This means that over 80% (48/59) of PHUs do not use the national utility as a primary electricity source. This mapping shows without a doubt that direct current–based renewable-energy alternatives may be better suited for targeting off-grid PHUs [19] (see [20] for technical differences between direct and alternating current electricity systems).

The internet distribution analysis showed that 82% (53/59) of all health facilities do not have official institutional-provided internet. However, 26% (19/72) had private internet across health facility types, as shown in Table 2. Three mobile telecommunications service providers (Africell, Orange, and SierraTel) provided internet in all health facility visits. Alternative internet networks such as fiber internet provided by the Ministry of Information [21] and satellite-based internet sources such as SpaceX can better serve off-the-grid health facilities [22]. In the event of no internet, an offline-first HIS solution will be most appropriate.

**Table 2.** Internet infrastructure based on the health facilities survey (N=72).

Health facility	No private internet, n	Has private internet, n
CHC <sup>a</sup>	19	6
MCHP <sup>b</sup>	14	1
CHP <sup>c</sup>	13	6
Hospital	7	6
Total	53	19

<sup>a</sup>CHC: Community Health Center.

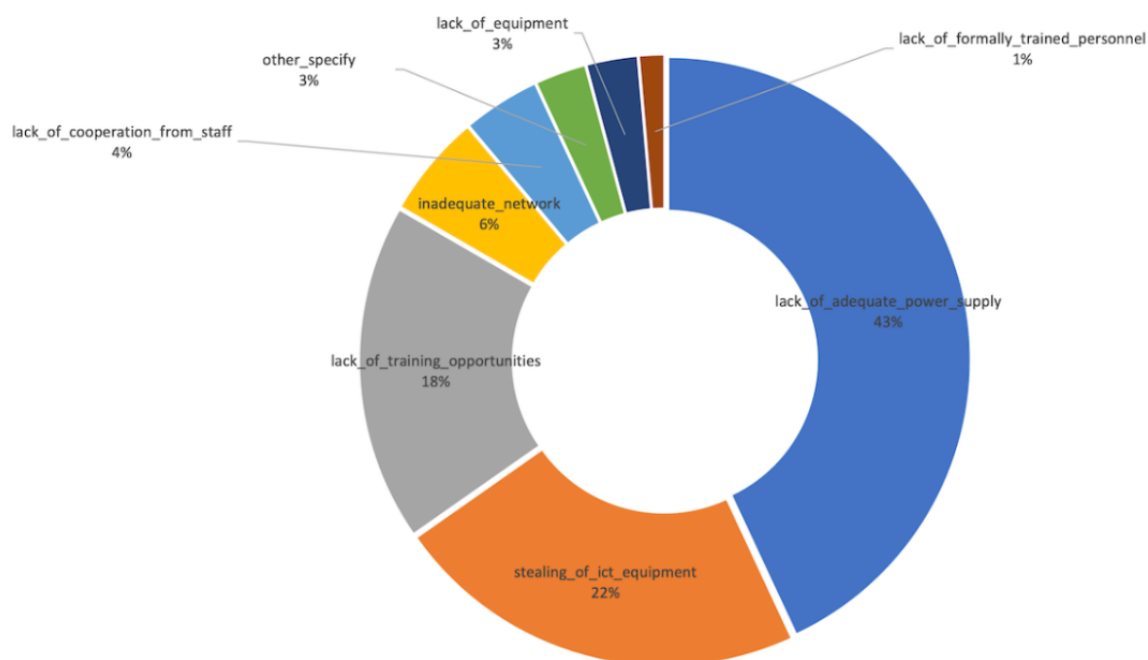
<sup>b</sup>MCHP: Maternal and Child Health Post.

<sup>c</sup>CHP: Community Health Post.

The majority of the PHUs surveyed had one tablet. Depending on the solutions deployed in these health facilities, one tablet per health facility may or may not be adequate [23]. Some advanced solutions may also not work on tablets, given that only 2 of the 59 PHUs had one computer each. Therefore, software solutions targeting these PHUs should be designed to be tablet-compliant. By contrast, hospitals did not all have the same number of computers. One hospital had 15 functional computers, another had 9, and six hospitals had 4 or fewer

functional computers. This shows that computing devices across the hospitals are not evenly distributed, indicating a significant computing infrastructure gap. Among the staff interviewed at the surveyed health facilities, 43% considered that lack of an adequate power supply was the biggest threat to digitization, as shown in Figure 5. Low ICT capacity and theft of digital devices were considered an equally significant threat to digitization in Sierra Leone.

**Figure 5.** The biggest threats to information communication technology efforts at the health facilities surveyed. ict: information and communications technology.



### Limitations

A key limitation of this survey is that the responses were self-reported, although we attempted mitigating against possible bias by ensuring that the enumerators sighted the infrastructure. However, some enumerators were not able to sight the infrastructure for several reasons.

### Conclusion

In this study, we explored the state of infrastructure enabling the digital health environment in Sierra Leone. We surveyed primary and secondary electricity sources, the type and nature

of computing hardware, and the internet and connectivity available at these health facilities. We aggregated how often these health facilities have electricity to help determine if a health facility information system can be viable. Disconnected PHUs or hospitals can use alternative electricity sources, fiber or satellite internet, and tablet hardware. This research will support the government in implementing strategies for bridging health facility infrastructure gaps. The next step from this study will be to extrapolate and determine the current infrastructure (electricity, internet, and computing hardware) costs from the national digital health cost plan [9].

### Acknowledgments

The Ministry of Health and Sanitation was supported for this study by United Nations Children's Fund (UNICEF) with support from United States Agency for International Development (USAID). The University of Malta also supported the manuscript publication fees.

### Authors' Contributions

FS is the approving authority for this study. EC designed and prepared the initial draft of this paper. EF and AK facilitated the data collectors' recruitment, training, field visits, and report reviews. RW and LG provided technical inputs and reviews of the manuscript.

### Conflicts of Interest

None declared.

### References

1. Stroetmann KA, Artmann J, Stroetmann V. Developing national eHealth infrastructures--results and lessons from Europe. *AMIA Annu Symp Proc* 2011;2011:1347-1354. [Medline: [22195196](#)]
2. Liu W, Park EK, Krieger U. eHealth interconnection infrastructure challenges and solutions overview. New York: IEEE; 2021 Oct 13 Presented at: 2012 IEEE 14th International Conference on e-Health Networking, Applications and Services



- (Healthcom); 2012; Beijing, China URL: <https://ieeexplore.ieee.org/document/6379417> [doi: [10.1109/HealthCom.2012.6379417](https://doi.org/10.1109/HealthCom.2012.6379417)]
3. Monitoring the building blocks of health systems: a handbook of indicators and their measurement strategies. World Health Organization. 2010. URL: [https://www.who.int/healthinfo/systems/WHO\\_MBHSS\\_2010\\_full\\_web.pdf](https://www.who.int/healthinfo/systems/WHO_MBHSS_2010_full_web.pdf) [accessed 2022-01-08]
  4. Omotosho A, Ayegba P, Emuoyibofarhe J, Meinel C. Current state of ICT in healthcare delivery in developing countries. *Int J Onl Eng* 2019 May 14;15(08):91. [doi: [10.3991/ijoe.v15i08.10294](https://doi.org/10.3991/ijoe.v15i08.10294)]
  5. Energy consumption in Sierra Leone. WorldData info. URL: <https://www.worlddata.info/africa/sierra-leone/energy-consumption.php> [accessed 2022-01-16]
  6. Herisse R, Herscowitz A. Power Africa in Sierra Leone. USAID. 2016 Sep 30. URL: [https://www.usaid.gov/sites/default/files/documents/1860/SierraLeoneCountryFactSheet.2016.09\\_FINAL.pdf](https://www.usaid.gov/sites/default/files/documents/1860/SierraLeoneCountryFactSheet.2016.09_FINAL.pdf) [accessed 2022-01-16]
  7. GSM Mobile Connectivity Index. GSMA. URL: <http://www.mobileconnectivityindex.com> [accessed 2020-05-12]
  8. Ministry of Health and Sanitation. Health facility availability. Open Data Sierra Leone. URL: <https://www.opendatasl.gov.sl/dataset/health-facility-availability> [accessed 2022-01-16]
  9. National Digital Health Strategy 2018-2023. Ministry of Health and Sanitation. 2018 Nov. URL: <https://mohs2017.files.wordpress.com/2019/02/sl-national-digital-health-strategy-nov-2018.pdf> [accessed 2021-11-11]
  10. Kirigia JM, Sambo LG, Renner A, Alemu W, Seasa S, Bah Y. Technical efficiency of primary health units in Kailahun and Kenema districts of Sierra Leone. *Int Arch Med* 2011 May 11;4(1):15 [FREE Full text] [doi: [10.1186/1755-7682-4-15](https://doi.org/10.1186/1755-7682-4-15)] [Medline: [21569339](https://pubmed.ncbi.nlm.nih.gov/21569339/)]
  11. District Health Information Systems (DHIS2). URL: <http://www.dhis2.org/> [accessed 2020-08-28]
  12. Report of the Sierra Leone National eHealth Coordination Hub Launch. Ministry of Health and Sanitation. 2018 Mar 30. URL: [https://mohs2017.files.wordpress.com/2017/12/report-of-the-ehealth-coordination-hub-launch\\_final.pdf](https://mohs2017.files.wordpress.com/2017/12/report-of-the-ehealth-coordination-hub-launch_final.pdf) [accessed 2021-01-20]
  13. National eHealth Strategy Toolkit. ITU. 2012. URL: <http://handle.itu.int/11.1002/pub/8069793a-en> [accessed 2022-01-19]
  14. Sample Size Calculator. Calculator.net. URL: <https://www.calculator.net/sample-size-calculator.html> [accessed 2019-01-18]
  15. Hunter JD. Matplotlib: A 2D Graphics Environment. *Comput Sci Eng* 2007 May;9(3):90-95. [doi: [10.1109/mcse.2007.55](https://doi.org/10.1109/mcse.2007.55)]
  16. Ngwenyama O, Morawczynski O. Factors affecting ICT expansion in emerging economies: An analysis of ICT infrastructure expansion in five Latin American countries. *Inf Technol Devel* 2010 Feb 19;15(4):237-258. [doi: [10.1002/itdj.20128](https://doi.org/10.1002/itdj.20128)]
  17. World Bank open data. World Bank. URL: <https://data.worldbank.org> [accessed 2021-01-22]
  18. Williams MDJ, Mayer R, Minges M. Africa's ICT infrastructure: building on the mobile revolution. World Bank. 2011. URL: <https://documents1.worldbank.org/curated/en/375261468204545149/pdf/632510PUB0ICT000ID0184540BOX361512B.pdf> [accessed 2020-11-12]
  19. Fathah A. Design of a boost converter. Bachelor of Technology Thesis. Department of Electrical Engineering National Institute of Technology Rourkela. 2013. URL: <http://ethesis.nitrkl.ac.in/4811/1/109EE0612.pdf> [accessed 2021-01-21]
  20. Liu X, Wang P, Loh PC. A hybrid AC/DC microgrid and its coordination control. *IEEE Trans Smart Grid* 2011 Jun;2(2):278-286. [doi: [10.1109/tsg.2011.2116162](https://doi.org/10.1109/tsg.2011.2116162)]
  21. Bundu BJ. Information Ministry, SALCAB Inspect National Fiber Backbone Phase 2 Project. Sierra Express Media. URL: <https://sierraexpressmedia.com/?p=88065> [accessed 2021-12-01]
  22. Foust J. SpaceX's space-Internet woes: Despite technical glitches, the company plans to launch the first of nearly 12,000 satellites in 2019. *IEEE Spectr* 2019 Jan;56(1):50-51. [doi: [10.1109/mspec.2019.8594798](https://doi.org/10.1109/mspec.2019.8594798)]
  23. Classification of digital health interventions v1.0. World Health Organization. 2018. URL: <https://www.who.int/reproductivehealth/publications/mhealth/classification-digital-health-interventions/en/> [accessed 2021-01-20]

## Abbreviations

- CHC:** Community Health Center  
**CHP:** Community Health Post  
**DPPI:** Directorate of Planning, Policy, and Information  
**HIS:** health information system  
**ICT:** information and communication technology  
**LMIC:** low- and middle-income country  
**MCH:** Maternal and Child Health  
**MCHP:** Maternal and Child Health Post  
**MoHS:** Ministry of Health and Sanitation  
**PHU:** Primary Health Care Unit  
**WHO:** World Health Organization

*Edited by C Lovis; submitted 29.04.21; peer-reviewed by P Nguyen, J Ye; comments to author 05.06.21; revised version received 16.06.21; accepted 26.07.21; published 03.02.22.*

*Please cite as:*

*Chukwu E, Garg L, Foday E, Konomanyi A, Wright R, Smart F*

*Electricity, Computing Hardware, and Internet Infrastructures in Health Facilities in Sierra Leone: Field Mapping Study*

*JMIR Med Inform 2022;10(2):e30040*

*URL: <https://medinform.jmir.org/2022/2/e30040>*

*doi: [10.2196/30040](https://doi.org/10.2196/30040)*

*PMID: [35113026](https://pubmed.ncbi.nlm.nih.gov/35113026/)*

©Emeka Chukwu, Lalit Garg, Edward Foday, Abdul Konomanyi, Royston Wright, Francis Smart. Originally published in JMIR Medical Informatics (<https://medinform.jmir.org>), 03.02.2022. This is an open-access article distributed under the terms of the Creative Commons Attribution License (<https://creativecommons.org/licenses/by/4.0/>), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Medical Informatics, is properly cited. The complete bibliographic information, a link to the original publication on <https://medinform.jmir.org/>, as well as this copyright and license information must be included.

Original Paper

# Disease Progression of Hypertrophic Cardiomyopathy: Modeling Using Machine Learning

Matej Pičulin<sup>1\*</sup>, PhD; Tim Smole<sup>1\*</sup>, MSc; Bojan Žunkovič<sup>1\*</sup>, PhD; Enja Kokalj<sup>1\*</sup>, MSc; Marko Robnik-Šikonja<sup>1\*</sup>, Prof Dr, PhD; Matjaž Kukar<sup>1\*</sup>, Prof Dr, PhD; Dimitrios I Fotiadis<sup>2\*</sup>, Prof Dr, PhD; Vasileios C Pezoulas<sup>2\*</sup>, MSc; Nikolaos S Tachos<sup>2\*</sup>, MSc; Fausto Barlocco<sup>3,4\*</sup>, PhD; Francesco Mazzarotto<sup>3,4,5\*</sup>, PhD; Dejana Popović<sup>6\*</sup>, MD, Prof Dr, PhD; Lars S Maier<sup>7\*</sup>, Prof Dr, PhD; Lazar Velicki<sup>8,9\*</sup>, MD, Prof Dr, PhD; Iacopo Olivotto<sup>3,4\*</sup>, MD, Prof Dr; Guy A MacGowan<sup>10\*</sup>, MD, PhD; Djordje G Jakovljević<sup>10,11\*</sup>, MD, Prof Dr, PhD; Nenad Filipović<sup>12\*</sup>, Prof Dr, PhD; Zoran Bosnić<sup>1\*</sup>, Prof Dr, PhD

<sup>1</sup>Faculty of Computer and Information Science, University of Ljubljana, Ljubljana, Slovenia

<sup>2</sup>Unit of Medical Technology and Intelligent Information Systems, Department of Materials Science and Engineering, University of Ioannina, Ioannina, Greece

<sup>3</sup>Department of Experimental and Clinical Medicine, University of Florence, Florence, Italy

<sup>4</sup>Cardiomyopathy Unit, Careggi University Hospital, University of Florence, Florence, Italy

<sup>5</sup>National Heart and Lung Institute, Imperial College London, London, United Kingdom

<sup>6</sup>Clinic for Cardiology, Clinical Center of Serbia, University of Belgrade, Belgrade, Serbia

<sup>7</sup>Department of Internal Medicine II (Cardiology, Pneumology, Intensive Care Medicine), University Hospital Regensburg, Regensburg, Germany

<sup>8</sup>Faculty of Medicine, University of Novi Sad, Novi Sad, Serbia

<sup>9</sup>Institute of Cardiovascular Diseases Vojvodina, Sremska Kamenica, Serbia

<sup>10</sup>Translational and Clinical Research Institute, Faculty of Medical Sciences, Newcastle University, Newcastle upon Tyne, United Kingdom

<sup>11</sup>Faculty of Health and Life Sciences, Coventry University, Coventry, United Kingdom

<sup>12</sup>Bioengineering Research and Development Center, Kragujevac, Serbia

\* all authors contributed equally

## Corresponding Author:

Matej Pičulin, PhD

Faculty of Computer and Information Science

University of Ljubljana

Večna pot 113

Ljubljana, 1000

Slovenia

Phone: 386 1 479 8226

Email: [matej.piculin@fri.uni-lj.si](mailto:matej.piculin@fri.uni-lj.si)

## Abstract

**Background:** Cardiovascular disorders in general are responsible for 30% of deaths worldwide. Among them, hypertrophic cardiomyopathy (HCM) is a genetic cardiac disease that is present in about 1 of 500 young adults and can cause sudden cardiac death (SCD).

**Objective:** Although the current state-of-the-art methods model the risk of SCD for patients, to the best of our knowledge, no methods are available for modeling the patient's clinical status up to 10 years ahead. In this paper, we propose a novel machine learning (ML)-based tool for predicting disease progression for patients diagnosed with HCM in terms of adverse remodeling of the heart during a 10-year period.

**Methods:** The method consisted of 6 predictive regression models that independently predict future values of 6 clinical characteristics: left atrial size, left atrial volume, left ventricular ejection fraction, New York Heart Association functional classification, left ventricular internal diastolic diameter, and left ventricular internal systolic diameter. We supplemented each prediction with the explanation that is generated using the Shapley additive explanation method.

**Results:** The final experiments showed that predictive error is lower on 5 of the 6 constructed models in comparison to experts (on average, by 0.34) or a consortium of experts (on average, by 0.22). The experiments revealed that semisupervised learning

and the artificial data from virtual patients help improve predictive accuracies. The best-performing random forest model improved  $R^2$  from 0.3 to 0.6.

**Conclusions:** By engaging medical experts to provide interpretation and validation of the results, we determined the models' favorable performance compared to the performance of experts for 5 of 6 targets.

(*JMIR Med Inform 2022;10(2):e30483*) doi:[10.2196/30483](https://doi.org/10.2196/30483)

## KEYWORDS

hypertrophic cardiomyopathy; disease progression; machine learning; artificial intelligence; AI; ML; cardiomyopathy; cardiovascular disease; sudden cardiac death; SCD; prediction; prediction model; validation

## Introduction

### Background

Recent reviews of machine learning (ML) applications in cardiovascular medicine [1,2] suggest that the use of ML is on the rise and that it is being adopted by doctors in their daily practice. ML applications in cardiology are reflected by augmenting medical practice by contributing to early diagnosis, risk stratification, and personalized therapeutics. The examples of such applications in other domains include modeling disease progression of Alzheimer disease [3,4], Parkinson disease [5], multiple sclerosis [6], chronic kidney disease [7], chronic liver disease [8], and others.

Cardiovascular disorders in general are responsible for 30% of deaths worldwide. Among them specifically, hypertrophic cardiomyopathy (HCM) is a genetic cardiac disease that is a cause of sudden cardiac death (SCD), especially among young adults and athletes [9]. Cardiovascular diseases represent groups of diseases that can greatly benefit from preemptive prediction, prevention, and proactive management; thus, this opens an opportunity for methods of artificial intelligence (AI) [2]. Disease progression is especially hard to detect in slow-progressing diseases, such as HCM, which is present in about 1 of 500 young adults [10]. Although HCM has 4 identified stages [11], patients with HCM can experience a sudden cardiac arrest or the disease can slowly progress over several years. Currently, the state-of-the-art HCM Risk-SCD calculator method for risk stratification of patients diagnosed with HCM [12] is widely used in practice. Although this method predicts the risk of SCD, no methods, to the best of our knowledge, are available for modeling the patient's clinical status up to 10 years ahead. Detection of cardiovascular risk for 10 years ahead is important and has been recently modeled for atherosclerotic cardiovascular disease [13].

In this paper, we propose a novel ML-based tool for predicting disease progression for patients diagnosed with HCM in terms of adverse remodeling of the heart during a 10-year period. The method consists of 6 contemporaneous predictive regression models that independently predict future values of the following 6 clinical characteristics: left atrial diameter (LA\_d), left atrial volume (LA\_Vol), left ventricular ejection fraction (LVEF), New York Heart Association (NYHA) functional classification, left ventricular internal diameter at end diastole (LVIDd), and left ventricular internal diameter at end systole (LVIDs). Each prediction is supplemented with the explanation that is generated using the Shapley additive explanation (SHAP) method [14].

Comparison between current and future values of these 6 parameters, as well as the interpretation of the change, generated by explanation methods, can help cardiologists gain insight into the disease progression trend for a given patient.

### Machine Learning Methods in Medicine

ML techniques are being frequently applied in medicine to improve the prediction of disease progression, extraction of medical knowledge for outcome research, therapy planning and support, and overall patient management [15]. A wide variety of ML approaches can solve challenging problems in these tasks. For example, diseases such as Alzheimer disease, diabetes, and chronic obstructive pulmonary disease (COPD) progress slowly over the years. For modeling of COPD, a Markov model was proposed by Wang et al [16], who also included a database of virtual patients. Their method successfully modeled progression trajectories, showing that multiple progression trajectories are possible for some diseases.

In cardiology, there are several works addressing disease progression trends related to different cardiological diseases. With the increase in computational power, ML has become a tool to analyze nonlinear dependencies that are present either in relational data or in images. Juarez-Orozco et al [17] emphasized the advantages of ML, especially deep learning, in cardiac nuclear imaging, where ML can aid with ischemia diagnosis and event prognosis. Sardar et al [18] emphasized the advantages of AI in interventional cardiology, which is promising to bring a paradigm shift in the practice of medicine by improving real-time clinical decision making and standardizing robotic medical procedures. While focusing on the use of ML in electrocardiogram (ECG) analysis, Elul et al [19] also stated the crucial disadvantages of ML, which include a lack of explanation, relating the automated diagnosis with medical knowledge, and transparency of the system's limitations. In their work, the authors proceeded to flag individual predictions that are irrelevant or not useful. To summarize, the mentioned works characterize AI as a developing tool that, with the synergy between humans and machines, will help transform medical practice and clinical care.

Further, a hybrid approach for progression of Parkinson disease [5] was successfully used by combining a variety of ML methods from different families: clustering, dimensional reduction, and incremental support vector regression. Deep learning was used for predicting Alzheimer disease, on average, about 6 years in advance [20] and for modeling Alzheimer disease progression [4]. Conditional restricted Boltzmann machines were also used for prediction of disease progression

[3]. The authors simulated patient trajectories using 18 months of longitudinal data of around 1900 patients and showed that patient-level simulations are feasible using ML and appropriate data.

Several other ML approaches also model disease progression well in other medical domains, such as kidney disease progression [7]. In this work, 9 ML approaches were tested: linear regression (LR), elastic net regression, lasso regression, ridge regression, support vector machines (SVMs), random forests (RFs), k-nearest neighbors (KNNs), neural networks (NNs), and XGBoost. Similarly, ML models were applied to the problem of disease progression for hepatitis C virus [8] for the 5-year prediction problem using longitudinal data. The authors' conclusion was that the boosted survival tree-based models using longitudinal data perform better than cross-sectional or linear models. Last but not least, ML was also used for disease progression and secondary progression detection for multiple sclerosis [6]. Several ML models were evaluated for predictions of disease severity in 6-10 years, such as KNNs, decision trees, LR, and SVMs. SVMs performed best.

To summarize, the overview indicates that ML models can be successfully applied to problems of predicting disease progression, which is also the goal of this paper. In the next subsection, we overview how ML approaches are used in cardiology, specifically for HCM, which is the focus of this paper.

### Machine Learning for Modeling Hypertrophic Cardiomyopathy

Most ML contributions to cardiovascular medicine focus on risk stratification of patients. One of the biggest obstacles to using data for a broader variety of ML applications is that data are usually stored in diverse repositories, which are not readily usable for cardiovascular research, due to various data quality challenges [2]. Where the data are readily available, different ML algorithms have been successfully used, such as Wasserstein generative adversarial networks [21], convolutional NNs [22,23], deep NNs [24], and boosted decision trees [25]. Some authors have tested multiple models, such as RFs, artificial NNs, SVMs, and Bayesian networks [26], or a combination of J48, naive Bayes, KNNs, SVMs, RFs, bagging, and boosting [27]. Cuocolo et al [1] overviewed ML methods in cardiology, emphasizing their successful applications for building clinical predictive models, for analyzing ECG signals and image data. For the latter problems, the most successful methods were NNs, deep NNs, and convolutional networks. Advances in prediction accuracy have also been made by using deep NNs to make predictions based on fast, large-scale genome-wide association studies [28].

HCM is a severe disease for which 4 stages of progression have been identified in the medical literature [11]. Current state-of-the-art ML mostly uses only statistical models, such as multivariate regression analysis, which uses preselected predictor variables of known medical importance. Cardiac magnetic resonance (CMR) images [29,30] and echocardiographic diagnostics [31] are found to be a good source

of important attributes for HCM identification. Recently, researchers have started proposing ML-based risk stratification for patients diagnosed with comorbidities to separate patients into low- and high-risk categories or several categories on a scale [32]. The medical literature is mostly focused on finding risk factors that identify increased risk of SCD in patients with HCM [12,33]. A study [34] presenting the guidelines used in risk stratification for patients with HCM proposed potential SCD modifiers. Maron et al [35] performed a similar study on older populations and also summarized risk factors that could prevent SCD. The continuation of this research [36] aimed to develop an accurate strategy to assess the reliability of SCD prediction methods in prevention of SCD in patients diagnosed with HCM.

It is important to note that patients with HCM who experience cardiac arrest are not identified by typical risk markers used in the American College of Cardiology or the statistical mathematical risk model by the European Society of Cardiology [37]. Therefore, new risk factors have been and still need to be considered and developed to provide additional information to better assess HCM risk. In our work, we focus on modeling the future development of HCM by predicting the change in relevant cardiac parameters 10 years ahead.

### Aims and Contributions

Novelties and contributions of this paper include:

- A disease progression system that comprises models for prediction of 6 contemporaneous relevant clinical parameters that are relevant to HCM for 10 years ahead. The system includes the implementation of the explanation methodology that provides interpretability of predictive models.
- Analysis of predictive performance if training data are extended using semisupervised learning or with artificial patient data.
- Validation of predictive accuracy with medical experts by comparing ML and human accuracy and by analyzing sensibility of the computer-generated prediction explanations.

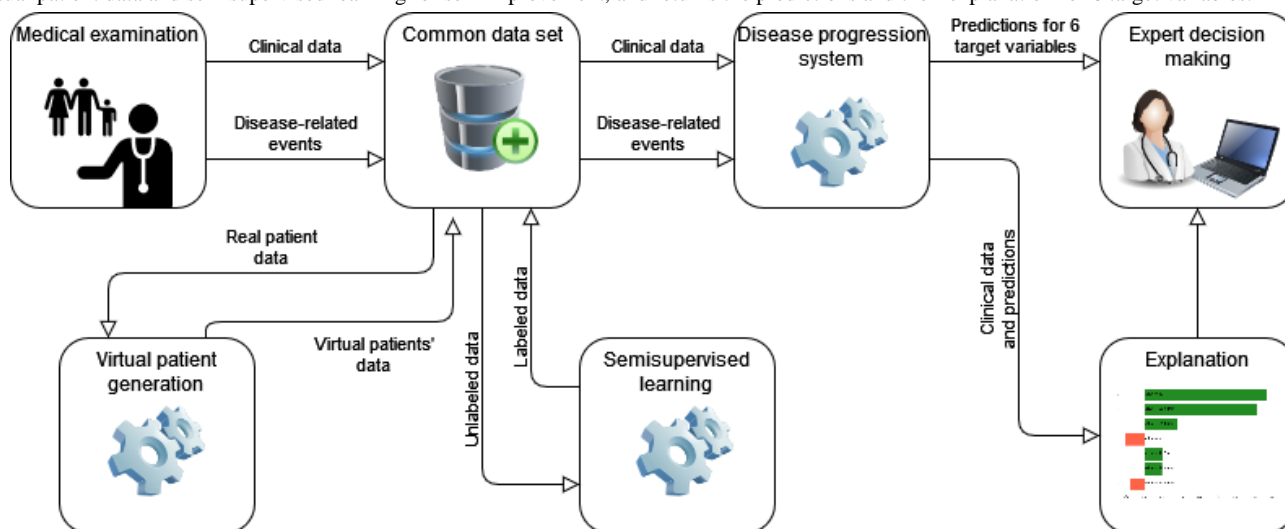
The aim of this paper is to develop a system capable of detecting slow progression of HCM based on longitudinal data.

## Methods

### Modeling Disease Progression

In this work, we modeled disease progression by predicting 6 relevant patient parameters 10 years in advance. These parameters are indicators of HCM and can be used to determine the stage of HCM according to the known guidelines [37]. Additionally, a preliminary analysis was performed to verify the prediction strength of the chosen parameters, validating our choice, as described in the Data Set section. The proposed disease progression system (Figure 1) takes as input patients' clinical data and data about their past disease-related events, such as dates of atrial fibrillation or syncope.

**Figure 1.** Overview of the proposed disease progression system. The system receives clinical data and disease-related events of a patient as input, uses virtual patient data and semisupervised learning for self-improvement, and returns the predictions and their explanation for 6 target variables.



The output of the system is a set of 6 contemporaneous target predictions for parameters:

- LA\_d
- LA\_Vol
- LVEF
- LVIDd
- LVIDs
- NYHA functional classification

In addition to predictions, the system also generates their explanations, revealing the factors with the largest impact on the increase or decrease in the 6 target variables throughout the 10-year period.

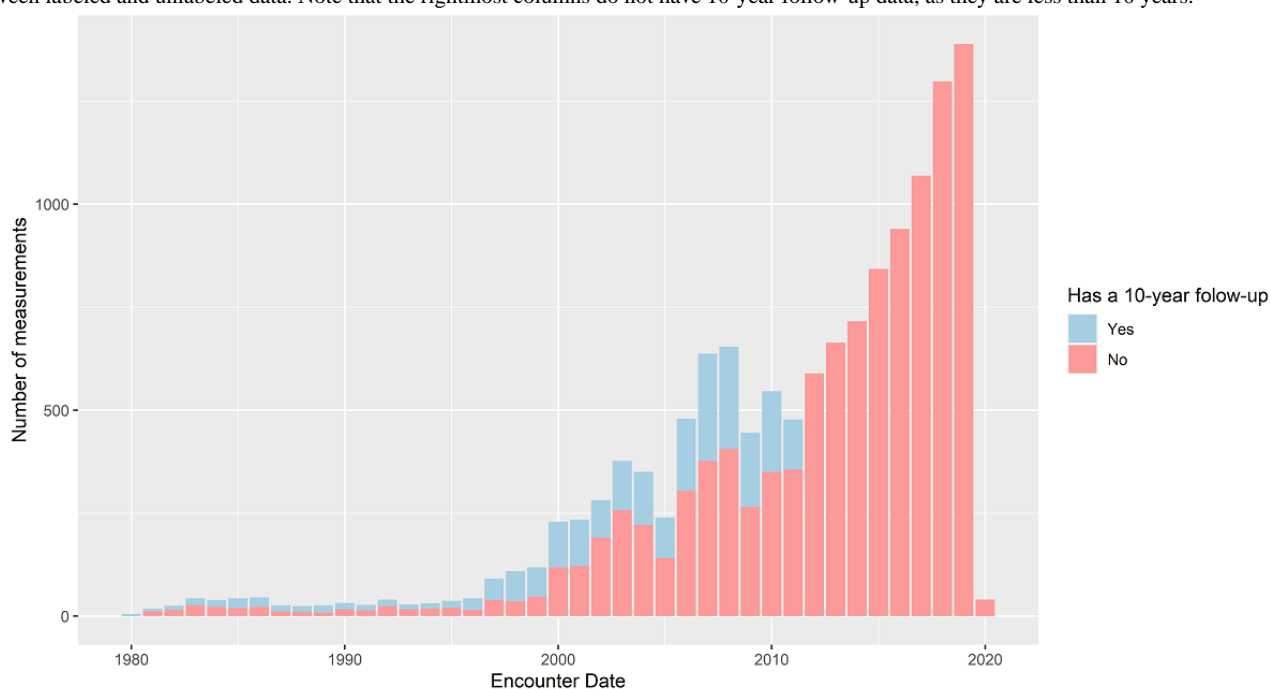
We trained the proposed disease progression system using supervised ML techniques. To further improve the results, we augmented the original data using unlabeled data (semisupervised learning) and virtual patients' data. We applied the semisupervised learning using patients without 10-year follow-ups and generated virtual patients' data using various techniques for artificial data generation. The semisupervised learning first predicted patients' targets using the trained models on labeled data, so they could be afterward included into the training data set. In the following subsections, we describe the data set, predictive modeling with supervised models, use of semisupervised learning and virtual patient data, and generation of prediction explanations.

### Data Set

The proposed approach was developed on a data set that was provided by the University of Florence as a result of its long-term clinical practice. The data set included patients who were enrolled over the past 40 years (Figure 2), and 1860 (80.24%) of 2318 patients had at least 1 available 10-year

follow-up. They were followed for an average duration of about 7 years and ranging up to 37 years. The data set contains longitudinal clinical data for 2318 patients diagnosed with HCM or patients that had a relative diagnosed with HCM (1457 [62.86%] male and 861 [37.14%] female patients). During the patients' visits, various clinical tests and relevant disease-related events were recorded. These data included general data (gender, age, height, weight, etc), genetic data (detected mutations), clinical tests (echocardiogram [echo], Holter monitoring, blood test, CMR, stress test), prescribed medications (type, start date, termination date), and disease-related events (eg, SCD, heart failure, transplant, abnormal Holter, pacemaker or implantable cardioverter defibrillator implantation). Echo was the leading diagnostic reference technique that was performed for the vast majority of patients and thus represents the main source of data. CMR was additionally used selectively due to its greater accuracy in measuring volumes. Although echo and CMR are treated separately and never computationally compared to each other in medical practice, we used CMR, where available, as an additional data modality to possibly improve prediction accuracy. In total, there were 6227 events recorded, of which 4902 (78.72%) events occurred in patients who were primarily diagnosed with HCM. The structure of the data set therefore allowed observing how patients' clinical characteristics change over time, which is essential for the desired modeling of HCM progression. The basic patient characteristics are shown in Table 1 for continuous parameters, Table 2 for binary parameters, and Table 3 for the remaining parameters. The characteristics were extracted from 10,318 measurements in total. Additionally, Table 4 shows the missing data numbers and percentages for the 6 selected target variables for their role as input or target variables.

**Figure 2.** Relationship between the amount of labeled and unlabeled data. The bars for Yes and No values are stacked, visually revealing the ratio between labeled and unlabeled data. Note that the rightmost columns do not have 10-year follow-up data, as they are less than 10 years.



**Table 1.** Basic characteristics of patients for basic continuous parameters (N=10,318).

Continuous parameter	Mean (SD)	Missing data, n (%)
Age (years)	52.1 (18.6)	4 (0.04)
Weight (kg)	73.4 (14.6)	2381 (23.08)
Height (cm)	169 (10.3)	2273 (22.03)
Body mass index (BMI)	25.6 (4.09)	2423 (23.48)
NYHA <sup>a</sup>	1.69 (0.73)	983 (9.53)

<sup>a</sup>NYHA: New York Heart Association.

**Table 2.** Basic characteristics of patients for basic binary parameters (N=10,318).

Binary parameter	1-value, n (%)	0-value, n (%)	Missing, n (%)
Alcohol	Yes, 103 (0.99)	No, 10,215 (99)	0
Drug	Yes, 18 (0.17)	No, 10,300 (99.83)	0
Smoking	Yes, 3437 (33.31)	No, 6881 (66.69)	0
Pregnancy	Yes, 443 (4.29)	No, 9875 (95.71)	2515 (24.37)
Gender	Male, 6400 (62.03)	Female, 3918 (37.97)	0

**Table 3.** Basic characteristics for groups of parameters (N=10,318)<sup>a</sup>.

Procedure	Parameters, n	Total missing values, n (%)
ECG <sup>b</sup>	9	45,839 (49.36)
Echo <sup>c</sup>	26	98,191 (36.60)
CMR <sup>d</sup>	10	81,174 (78.67)

<sup>a</sup>The table shows aggregated statistics for several parameters obtained from the same procedure. The percentage for each procedure is obtained as follows: [Total missing values/(Parameter × N)] × 100.

<sup>b</sup>ECG: electrocardiogram.

<sup>c</sup>Echo: echocardiogram.

<sup>d</sup>CMR: cardiovascular magnetic resonance.

**Table 4.** Absolute number and percentage of missing values of target variables as class and as input (N=10,318).

	LA_d <sup>a</sup> , n (%)	LVEF <sup>b</sup> , n (%)	NYHA <sup>c</sup> , n (%)	LVIDd <sup>d</sup> , n (%)	LVIDs <sup>e</sup> , n (%)	LA_Vol <sup>f</sup> , n (%)
Target	8569 (83.05)	8481 (82.19)	8313 (80.57)	8607 (83.42)	9336 (90.48)	8631 (83.65)
Input	2691 (26.08)	2399 (23.25)	983 (9.53)	2517 (24.39)	5329 (51.65)	3680 (35.67)

<sup>a</sup>LA\_d: left atrial diameter.

<sup>b</sup>LVEF: left ventricular ejection fraction.

<sup>c</sup>NYHA: New York Heart Association.

<sup>d</sup>LVIDd: left ventricular internal diameter at end diastole.

<sup>e</sup>LVIDs: left ventricular internal diameter at end systole.

<sup>f</sup>LA\_Vol: left atrial volume.

First, we transformed the available data set into a suitable form for predicting a 10-year change in relevant parameters using ML. Similarly, in other real-world data sets, most of the clinical tests were missing many patients or measurements were not taken for the whole span of 10 years (Figure 2). To address this issue, we preprocessed the data as follows:

- Formation of training examples: Since not all clinical tests can be conducted on the same day or in the same month, we defined a training example as a set of measurements within a time frame of 1 year. Such time frame corresponds to the annual regular visit period of patients and allows enough time for relevant changes in the observed parameters to become noticeable, as the disease slowly progresses. If the patient had a certain test performed multiple times within this time frame, multiple tests were treated as separate measurements. In case a certain type of test was not performed in the 1-year time frame, the corresponding variables were recorded as missing. Constructing training examples in this way yielded a data set with 13,386 examples, with 3.9 (SD 4.8) examples per patient.
- Imputation of missing data: The missing values in the data set, either because of nonperformed tests or because of erroneous input of data, were imputed by copying the closest past values (sensible because the progression of HCM is slow; used on numerical and categorical attributes), imputing values of a healthy patient (sampled from the normal distribution; used for numerical attributes), or

imputing mean values where healthy values were unknown (used on numerical and categorical attributes). Since measurements were not taken at equidistant time intervals, we used linear interpolation for computing equidistant measurement approximations.

We used the formed training examples as input to supervised learning algorithms. Prior to modeling, we evaluated the quality of attributes, which is important for decreasing learning complexity, avoiding overfitting, and, therefore, improving the simplicity and performance of ML methods. To facilitate learning with NNs, we also scaled the values to the interval [0,1] and encoded nominal values using the one-hot encoding method.

We used RReliefF [38], adaptation of the ReliefF feature selection algorithm, for regression problems. RReliefF calculates how well a feature's values distinguish between distant labels of instances that are close to each other and considers feature interactions. We selected 21 (18.7%) of 112 attributes based on the average rank across all 6 target variables for further supervised learning. Feature scores for 21 selected features are shown in Table 5, along with their average ranks across 6 trained predictive models. After removing highly correlated features (eg, the weight feature that correlates to the body surface area [BSA] and height), the final set of attributes contained all target variables (regardless of their rank) and the best-performing attributes, based on average rank.



**Table 5.** Selected attributes using RReliefF.<sup>a</sup>

Variable <sup>b</sup>	LA_d <sup>c</sup> score	LVEF <sup>d</sup> score	NYHA <sup>e</sup> score	LVIDd <sup>f</sup> score	LVIDs <sup>g</sup> score	LA_Vol <sup>h</sup> score	Average rank
<b>Anthropometric parameters</b>							
<i>Age</i>	0.198	0.194	0.166	0.142	0.166	0.158	1.000
<i>Gender</i>	0.051	0.037	0.043	0.055	0.058	0.022	12.500
<i>Height</i>	0.057	0.064	0.045	0.075	0.051	0.029	9.167
<i>BSA</i> <sup>i</sup>	0.075	0.073	0.053	0.095	0.085	0.045	4.167
<b>Risk factors</b>							
<i>Smoking</i>	0.063	0.046	0.052	0.032	0.069	0.082	7.500
<i>Presence of hypercholesterolemia</i>	0.072	0.042	0.052	0.039	0.044	0.056	9.667
History of syncope	0.026	0.036	0.029	0.022	0.029	0.048	20.000
<i>Family history of HCM</i> <sup>j</sup>	0.056	0.060	0.061	0.047	0.052	0.066	5.833
Family history of SCD <sup>k</sup>	0.027	0.051	0.032	0.031	0.051	0.049	14.667
<b>Clinical, ECG<sup>l</sup>, and echo<sup>m</sup> parameters</b>							
NYHA	0.011	0.017	0.069	0.007	0.027	0.022	33.000
Presence of atrial fibrillation	0.055	0.036	0.048	0.018	0.026	0.068	16.333
QRS duration	0.035	0.046	0.029	0.039	0.026	0.039	17.167
<i>Interventricular septum ( IVS)</i>	0.043	0.052	0.049	0.041	0.057	0.052	8.167
LA_d	0.078	0.037	0.036	0.018	0.031	0.070	15.000
LA_Vol	0.055	0.029	0.026	0.012	0.025	0.059	24.000
LVIDs	0.017	0.022	0.027	0.029	0.043	0.031	25.167
LVIDd	0.021	0.017	0.017	0.036	0.044	0.026	27.667
LVEF	0.018	0.051	0.019	0.014	0.050	0.013	27.833
<b>Genetics</b>							
<i>Mutation MYBPC3</i>	0.045	0.041	0.039	0.051	0.052	0.059	9.667
<i>Mutation MYH7</i>	0.037	0.044	0.034	0.040	0.066	0.023	14.667
Negative genetics	0.036	0.037	0.027	0.043	0.030	0.031	18.667

<sup>a</sup>The table shows RReliefF feature scores and the average ranks for each target variable.

<sup>b</sup>Names of the 10 highest-ranked variables are italicized.

<sup>c</sup>LA\_d: left atrial diameter.

<sup>d</sup>LVEF: left ventricular ejection fraction.

<sup>e</sup>NYHA: New York Heart Association.

<sup>f</sup>LVIDd: left ventricular internal diameter at end diastole.

<sup>g</sup>LVIDs: left ventricular internal diameter at end systole.

<sup>h</sup>LA\_Vol: left atrial volume.

<sup>i</sup>BSA: body surface area.

<sup>j</sup>HCM: hypertrophic cardiomyopathy.

<sup>k</sup>SCD: sudden cardiac death.

<sup>l</sup>ECG: electrocardiogram.

<sup>m</sup>Echo: echocardiogram.

## Predictive Modeling With Supervised and Semisupervised Machine Learning

To model the relationship between input patient data and target variables, we applied the following supervised learning algorithms:

- RFs [39,40] are an ensemble prediction model that construct multiple randomized decision trees. The implementations of an RF classifier in the R statistical package (library ranger) and the Python Scikit-Learn package [41] were used.

Each forest used between 500 and 1500 trees, and the Gini index was used as the attribute-splitting rule.

- Gradient boosting (XGBoost) [42] is an ensemble of weak decision tree predictors, implemented in the open source software library XGBoost.
- LR is a traditional method of finding a linear dependence between attributes and the selected target variable.
- NNs mimic the architecture and working of brain neurons. We used 1 input and 1 output layer and 1 or several hidden layers. In the optimization process, we optimized several learning parameters, such as the learning rate, number of hidden layers, sizes of layers, regularization, sample weights, class weights, dropout, and batch normalization.

The best hyperparameters of these algorithms were tuned using Bayesian optimization and random search implemented in *keras-tuner* [43].

### Semisupervised Learning and Virtual Patients

Semisupervised learning is increasingly used in medicine, especially for medical image segmentation [44-46]. This approach allows labeling a large amount of unlabeled data using only a small portion of labeled data. The majority (ie, 83.9% averaged over 6 target variables) of patients' data did not have records for the follow-up after 10 years. These unlabeled data were used as examples for semisupervised learning, producing a teacher model. The unlabeled examples were labeled with the supervised learning predictive model (see the Predictive Modeling With Supervised and Semisupervised Machine Learning section) and added to the training set. After that, a new model (also called a student model) was trained and kept if it achieved better performance on the test set than the teacher model.

To further improve the results of semisupervised learning, we used artificially generated data (ie, virtual patients). Virtual data generation can sometimes replace experiments in biomedical experiments on animals [47]. Specifically in cardiovascular modeling, patient-specific virtual patient modeling has recently made major progress in improving diagnoses [48]. We evaluated the performance and appropriateness of several virtual patient data generators for this task, such as the generator based on the multivariate normal and log-normal distribution (MVND and log-MVND) [49], and nonparametric methods using supervised tree ensembles, unsupervised tree ensembles, radial basis function-based NNs [50], and Bayesian networks [51]. As the final data generator, we chose the unsupervised tree ensembles, which exhibited the highest level of agreement between the real and the virtual distributions, computed with the Kolmogorov-Smirnoff goodness-of-fit statistical test [52]. We generated 10,000 virtual patient examples, with 20 most important features, listed in Data Set section.

### Explanation of the Predictive Model

Supervised ML models often exhibit a black-box nature, meaning that they can model data but not provide an explanation for the contained knowledge as well as the reasoning used in predictions. This means that the models lack transparency and interpretability. To address this, explanation methods provide justification for each prediction and assess features with the

highest impact [53]. This is important in risk-sensitive ML application areas, such as medicine, where the predictions of ML models need to be understood as they may represent a basis for further medical interventions.

In our work, we applied the SHAP method [14], which is a model-agnostic method, generating an explanation for different ML models in a unified form. The method uses theoretically sound concepts of Shapley values from cooperative game theory for computing contributions of each individual attribute value and of each attribute overall. The generated explanations visualize the most relevant attributes that contribute to higher or lower prediction values. The explanations can be computed either for a single patient's predictions or summarized over all patients to discover more general relationships between attributes and the model's predictions.

## Results

### Models' Comparison

To evaluate and compare the performance of the 6 predictive models, we used stratified 10-fold cross-validation. For each of the 6 predictive problems, 4 different regression models were evaluated (LR, RF, gradient-boosted [GB] trees, and NN). The following parameters were varied in tests:

- Application of semisupervised learning (denoted with S)
- Addition of virtual patients' data into the learning data set (denoted with VP)
- Use of all 112 features (denoted with All) or only a subset of the 21 best features (denoted with Subset)
- Interpolation of data points so that measurements were equidistant (denoted with I)

In all, 28 different combinations of the parameters were used in experiments. Some combinations were omitted due to limitations (eg, VP generators cannot generate data for all 112 attributes, so VP was evaluated only with the subset of attributes) or excessive time complexity (eg, the use of virtual patients with NNs).

### Performance of Predictive Models

To compare the accuracy of the obtained models, we computed the following 4 metrics: mean absolute error (MAE), root-mean-square error (RMSE), and 2 variations of the relative root-mean-square error (RRMSE<sub>mean</sub> and RRMSE<sub>const</sub>). The MAE measures the average absolute difference between predicted and true values over all examples in the test set. The RMSE addresses the issue that the squared values of the MSE are hard to interpret. The RRMSE measures the relative ratio between the obtained model and the baseline model. We computed 2 variations of the RRMSE with 2 different baseline models: mean predictor and constant predictor. With the RRMSE<sub>mean</sub>, we compared the performance of the obtained model to the model that returned the mean of the target variable over all patients (mean predictor), while with the RRMSE<sub>const</sub>, we compared the obtained model to the model that assumed that the value of the target variable would remain constant/unchanged over the 10-year period (constant predictor).

We summarized (Table 6) the performance of the best-performing predictive models (RF, LR, GB, NN) and parameters (S, VP, All/Subset) for each target variable. We could see that the top-performing regression models were the RF and the GB tree for all target variables. We achieved the best results by applying semisupervised learning (S) for all target variables and using virtual patients (VP) for 5 of 6 target variables. For all targets, the best results were obtained by learning from a subset of the 21 most important features. The values of both RRMSE metrics revealed that the model performs better than the baseline models (their values are less than 1.0), with the model for the LA\_d target achieving the lowest predictive error.

To further evaluate the contribution of different data augmentation strategies, we compared the results on different patient sets: original (All features), subset of best features (Subset), virtual patients (VP), semisupervised learning (S), and the combination of the latter 2 (S + VP). The obtained results, shown for the best-performing RF model, are given in Figure 3, which compares the  $R^2$  metrics for each individual target parameter. The additional detailed results for the other models are given in Multimedia Appendix 1. The obtained results reveal the benefits of reducing the feature space, as well as applying the used data augmentation methods.

In the following subsection, we apply the explanation methodology that helps interpret the computed predictions and their contributing feature values.

**Table 6.** Comparison of the best-performing models for each target variable.

Target	Model and parameter	MAE <sup>a</sup>	RMSE <sup>b</sup>	RRMSE <sup>c</sup> <sub>mean</sub>	RRMSE <sub>const</sub>
LA_d <sup>d</sup>	RF <sup>e</sup> : S <sup>f</sup> +VP <sup>g</sup> +Subset	3.4	4.73	0.54	0.46
LA_Vol <sup>h</sup>	RF: S+VP+Subset	18.4	26.73	0.56	0.47
LVEF <sup>i</sup>	GB <sup>j</sup> : S+Subset	4.92	6.73	0.67	0.61
LVIDd <sup>k</sup>	RF: S+VP+Subset	3.53	5.26	0.68	0.64
LVIDs <sup>l</sup>	RF: S+VP+Subset	3.42	4.81	0.66	0.56
NYHA <sup>m</sup>	RF: S+VP+Subset	0.39	0.5	0.67	0.66

<sup>a</sup>MAE: mean absolute error.

<sup>b</sup>RMSE: root-mean-square error.

<sup>c</sup>RRMSE: relative root-mean-square error.

<sup>d</sup>LA\_d: left atrial diameter.

<sup>e</sup>RF: random forest.

<sup>f</sup>S: application of semisupervised learning.

<sup>g</sup>VP: addition of virtual patients' data into the learning data set.

<sup>h</sup>LA\_Vol: left atrial volume.

<sup>i</sup>LVEF: left ventricular ejection fraction.

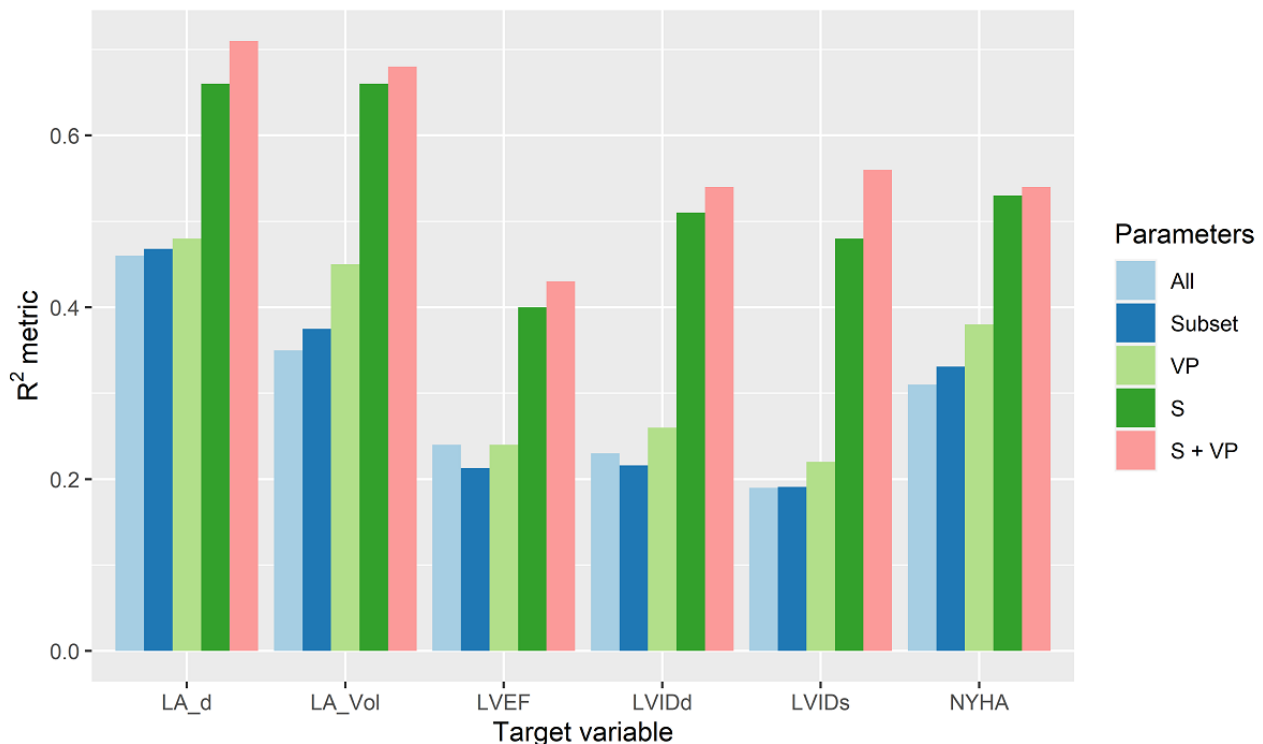
<sup>j</sup>GB: gradient boosted.

<sup>k</sup>LVIDd: left ventricular internal diameter at end diastole.

<sup>l</sup>LVIDs: left ventricular internal diameter at end systole.

<sup>m</sup>NYHA: New York Heart Association.

**Figure 3.** Plotted results for the  $R^2$  statistic for each target variable using different sets (input parameters). Note that VP, S, and S + VP are used on feature subsets. LA\_d: left atrial diameter; LA\_Vol: left atrial volume; LVEF: left ventricular ejection fraction; LVIDd: left ventricular internal diameter at end diastole; LVIDs: left ventricular internal diameter at end systole; NYHA: New York Heart Association; S: application of semisupervised learning; VP: addition of virtual patients' data into the learning data set.



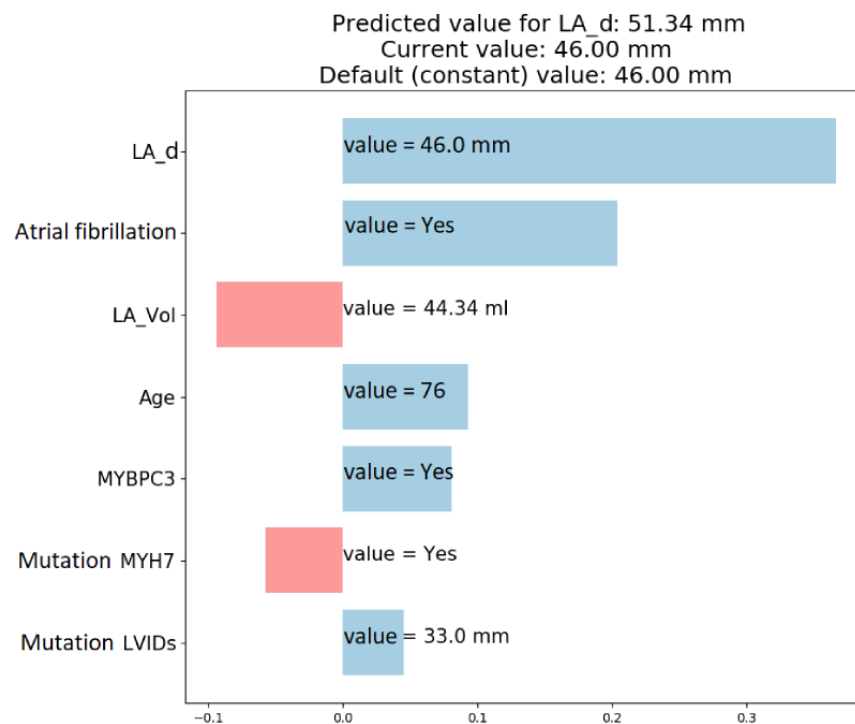
### Explanation of Predictions

To augment the output of prediction models, we applied the SHAP method [14] for computing explanations of individual predictions. The explanation of a single prediction consists of relevant textual, graphical, and numerical data that allows understanding of the relationships between the features of the patient and the model's prediction. It also consists of a list of the most relevant features that influence the prediction, along with their contribution values that define whether the feature value either supports the predicted value or opposes it. The direction of the impact (ie, sign of the contribution value) is denoted using different colors.

An example of the explanation generated for the prediction for the target LA\_d (Figure 4) is presented here. Features' contributions are sorted in descending order, and the graph contains only the features for which the sum of their

contributions reflects 95% of the difference between the initial parameter value and the predicted value after 10 years. The green and red bars thus denote positive and negative contributions of the impact for individual feature values, respectively, showing the factors contributing to the increase or decrease in the LA\_d value. We can see that the LA\_d, atrial fibrillation, age, mutation MYBPC3, and LVIDs features contributed to the increase in the predicted value for LA\_d over time, while the LA\_Vol and mutation MYH7 features contributed to the decrease in the predicted value for LA\_d. Because the overall increasing impact was more prominent, the final predicted value (51.34) was higher than the baseline prediction, which is also the current patients' value of LA\_d (46.00). Larger magnitudes of the features' contributions correspond to larger changes in the prediction value. For example, LA\_d contributed the most (approximately 30%) to the increase in the predicted value.

**Figure 4.** Example of an explanation of the prediction for the target variable LA\_d. LA\_d: left atrial diameter; LA\_Vol: left atrial volume; LVIDs: left ventricular internal diameter at end systole.



### Validation With Medical Experts

Besides evaluation of prediction models with statistical measures conducted in 2 previous sections, we engaged medical experts to provide further interpretation and validation of the results. First, we compared the accuracy of predictive models with the accuracy of human experts, which was obtained by using a survey (Multimedia Appendix 2). Second, we checked whether prediction explanations were sensible and consistent with the experts' medical knowledge about HCM.

We prepared a questionnaire for medical experts and distributed it to several medical universities and cardiology clinics. The questionnaire included data about complete medical cases (measurements, events, and medication data) for 10 patients, and the experts were asked to study them and complete the following 2 tasks:

- Predict the magnitude of the 10-year change in the 6 studied clinical parameters (LA\_d, LA\_Vol, LVEF, LVIDd, LVIDs, and NYHA) and mark it on a discrete scale from  $-3$  to  $3$ , where  $-3$  and  $3$  represented the biggest-possible decrease and increase, respectively. Possible magnitudes of change were represented using discrete intervals, as the prediction of an exact value is a difficult task that does not take place in medical practice.
- Evaluate whether the statements generated from the explanation (eg, “The current value of parameter LA\_d will cause a decrease in LA\_d”) are true or false. For each patient, 6 such statements were generated, covering the features with the highest contribution. More specifically, the questionnaire included evaluation questions for 6

parameters that contribute to a change in LA\_d, 4 for LA\_Vol, 5 for LVEF, 6 for LVIDd, 7 for LVIDs, and 4 for NYHA.

The questionnaire was fully completed by 13 experts with 16 (SD 8) years of experience. In the following subsections, we present the analysis of the answers.

### Validation of Prediction Accuracy

To compare the prediction accuracy between the experts and the ML model, we first discretized the model's predictions into discrete intervals so that they could be compared to the discrete intervals, predicted by the experts. We performed the discretization using bins of width  $0.25\sigma$ , where  $\sigma$  is the SD of the variable. Further, we calculated the following prediction errors:

- Mean prediction error of the discretized model prediction (denoted with MD)
- Mean prediction error made by individual medical experts (denoted with E)
- Mean prediction error of the consortium prediction (ie, the average prediction of all doctors, denoted with C)

We could see that the mean prediction error of the discretized model MD (Table 7) was the lowest for all target variables except for LA\_d. The mean errors of consortium predictions C were lower than the predictions of individual experts for all parameters, which indicates that the mutual consolidation of different doctors' opinions reduced the error of their joint predictions. The consortium prediction error also turned out to be the lowest for the parameter LA\_d and thus better than the error of the ML model.

**Table 7.** Mean absolute error (MAE) of the discretized model predictions (MD), individual experts (E), and the entire consortium (C).

Target/prediction	Model (MD), MAE (SD)	Expert (E), MAE (SD)	Consortium (C), MAE (SD)
NYHA <sup>a</sup>	<i>0.30 (0.48)</i> <sup>b</sup>	0.84 (0.69)	0.56 (0.34)
LA_d <sup>c</sup>	1.70 (0.82)	1.69 (0.97)	<i>1.66 (0.70)</i> <sup>b</sup>
LA_Vol <sup>d</sup>	<i>1.00 (0.82)</i> <sup>b</sup>	1.25 (0.98)	1.13 (0.63)
LVIDd <sup>e</sup>	<i>0.80 (0.63)</i> <sup>b</sup>	1.09 (0.91)	1.00 (0.77)
LVIDs <sup>f</sup>	<i>0.50 (0.71)</i> <sup>b</sup>	1.02 (0.86)	0.88 (0.68)
LVEF <sup>g</sup>	<i>0.90 (0.88)</i> <sup>b</sup>	1.32 (0.90)	1.28 (0.79)

<sup>a</sup>NYHA: New York Heart Association.

<sup>b</sup>The lowest achieved errors are italicized.

<sup>c</sup>LA\_d: left atrial diameter.

<sup>d</sup>LA\_Vol: left atrial volume.

<sup>e</sup>LVIDd: left ventricular internal diameter at end diastole.

<sup>f</sup>LVIDs: left ventricular internal diameter at end systole.

<sup>g</sup>LVEF: left ventricular ejection fraction.

### Validation of the Model Explanation

To validate the generated model explanations, we analyzed the agreement of experts with statements generated about the features' influence in 2 steps. First, we calculated the agreement ratio for individual features that were included in the questionnaire, grouped by each of the 6 target variables. Second, we calculated the overall agreement of experts with the explanation for each of the 6 target parameters, based on the agreement data about all features that contributed to their prediction.

The results (Table 8) of the analysis provided the ratio of agreement between different parameters for each target variable, as well as their overall agreement. The highest agreement ratio was achieved for target attributes NYHA (1.00), LA\_Vol (0.75), and LVIDd (0.67). The last column (Average agreement) summarizes the results across all used features. The results, in decreasing order, of the last column show that the majority of the experts agreed, especially with the explanations for the targets NYHA (average agreement of 0.73) and LVIDd (average agreement of 0.52). By comparing Tables 7 and 8, we consistently see that the experts least agreed with explanations for the target LA\_d, for which the predictive model achieved a

larger error than individual experts or the entire consortium. In cases where the predictive model achieved better predictive accuracy than the experts (Table 7) and the agreement of the experts with the explanation was lower (Table 8), for example, for LVEF, LA\_Vol, and LVIDs, there are 3 possible explanations:

- The generated explanation might, indeed, provide incorrect information.
- The generated explanation might explain novel relationships between features and target parameters that have not been observed or documented so far.
- It was hard for the experts to evaluate the claims in the questionnaire about the influence of particular features, as these tasks deviate from the established medical practice and require the experts to rely on their subjective experience.

For establishing the reasons for imperfect agreement between the explanation and the experts, further investigation is therefore required. We can conclude that the results provide some evidence that the generated prediction explanation might provide a complementary view at the prediction of HCM-related parameters. Such explanations might represent a tool that the experts could consult while making their decisions.

**Table 8.** Agreement ratios between experts and prediction explanations for parameters that contribute to predicting each target variable. The last two columns provide summary statistics.

Target variable and parameters	Expert agreement	Summary Ratio of agreed features from at least 50% of experts, n	Average agreement, n
<b>NYHA<sup>a</sup></b>			
<i>LA_d<sup>b</sup></i>	0.77 <sup>c</sup>		
<i>Age</i>	0.77 <sup>c</sup>	1.00 (4/4)	0.73
<i>LA_Vol<sup>d</sup></i>	0.62 <sup>c</sup>		
<i>Atrial fibrillation</i>	0.77 <sup>c</sup>		
<b>LVIDd<sup>e</sup></b>			
<i>BSA<sup>f</sup></i>	0.15		
<i>Gender</i>	0.85 <sup>c</sup>		
<i>LVIDd</i>	0.65 <sup>c</sup>	0.67 (4/6)	0.52
<i>QRS duration</i>	0.69 <sup>c</sup>		
<i>LVEF<sup>g</sup></i>	0.23		
<i>Mutation MYH7</i>	0.54 <sup>c</sup>		
<b>LVEF</b>			
<i>QRS duration</i>	0.38		
<i>Presence of hypercholesterolemia</i>	0.54 <sup>c</sup>		
<i>Syncope</i>	0.46	0.40 (2/5)	0.49
<i>Gene_Testing_Performed</i>	0.69 <sup>c</sup>		
<i>NYHA</i>	0.38		
<b>LA_Vol</b>			
<i>LA_Vol</i>	0.69 <sup>c</sup>		
<i>BSA</i>	0.54 <sup>c</sup>	0.75 (3/4)	0.48
<i>Age</i>	0.15		
<i>Atrial fibrillation</i>	0.54 <sup>c</sup>		
<b>LVIDs</b>			
<i>LA_d</i>	0.38		
<i>LVIDd</i>	0.38		
<i>LA_Vol</i>	0.62 <sup>c</sup>		
<i>BSA</i>	0.85 <sup>c</sup>	0.43 (3/7)	0.47
<i>Mutation MYBPC3</i>	0.62 <sup>c</sup>		
<i>Interventricular septum (IVS)</i>	0.38		
<i>Family history of HCM<sup>h</sup></i>	0.08		
<b>LA_d</b>			
<i>LA_d</i>	0.85 <sup>c</sup>		
<i>Atrial fibrillation</i>	0.15		
<i>BSA</i>	0.08	0.17 (1/6)	0.36
<i>IVS</i>	0.38		

Target variable and parameters	Expert agreement	Summary
		Ratio of agreed features from at least 50% of experts, n
		Average agreement, n
Age	0.31	
LVEF	0.38	

<sup>a</sup>NYHA: New York Heart Association.

<sup>b</sup>LA\_d: left atrial diameter.

<sup>c</sup>Names of parameters with agreement higher than 50% are italicized.

<sup>d</sup>LA\_Vol: left atrial volume.

<sup>e</sup>LVIDd: left ventricular internal diameter at end diastole.

<sup>f</sup>BSA: body surface area.

<sup>g</sup>LVEF: left ventricular ejection fraction.

<sup>h</sup>HCM: hypertrophic cardiomyopathy.

## Discussion

### Principal Results

We presented a disease progression system for patients diagnosed with HCM that is based on predicting 6 target parameters (LA\_d, LA\_Vol, LVIDd, LVIDs, LVEF, and NYHA) for 10 years ahead using supervised ML models. The experiments revealed good ML performance for all targets, with the achieved predictive error lower than the error of the default predictors. The experiments also revealed that semisupervised learning and the artificial data from virtual patients helped achieve even higher predictive accuracy for all 6 targets. Finally, we validated our approach with human experts using a structured questionnaire and determined the models' favorable performance compared to performance of experts for 5 of 6 targets.

### Limitations

The design of the study carried several limitations, stemming from the fact that this work was based on real-world data that are expensive to obtain and are subject to noise. The first limitation of this study is that it was based only on a single medical center data set. To further validate this study, it would be beneficial to independently evaluate the models with data sets from other centers or extend the existing data set with more data. Additionally, the benefit for including more data could also be in diminishing a potential bias of our data set, which could potentially include a population distribution that is different from other medical centers and thus different ranges of recorded parameters, which we did, in fact, observe in some cases. Additionally, in the perfect but rather unrealistic scenario due to its cost, both data modalities (echo and CMR) would be available for all patients, which would allow us to use the CMR data as an additional data source for all patients. Due to the unavailability of such data at the time of the study or data that were structured differently, we leave this for our further work.

Further, to prepare the data to be used for ML and obtain stable predictions, we used several preprocessing and data augmentation steps. Since we are dealing with real medical data, this opens questions of how different data transformations influence our predictions. Hence, a sensitivity study of the

results would be required, as well as determining how the patient's record time frame and predicted risk time frame influence the achieved accuracies. An additional limitation of the performed validation was that the ML results were compared to the inputs of medical experts in the structured survey instead of their free diagnoses and evaluations. Although this was required to unify the structure of human answers to enable statistical comparisons, the form of survey might introduce its own bias.

The described limitations, along with our further research questions and ideas, open several ideas for future study directions. First, we will evaluate the proposed system on an independent cardiological data set (eg, the Sarcomeric Human Cardiomyopathy Registry [SHaRe]) [54]. Second, as our current approach provides future predictions for 6 independent parameters, the outputs will be further combined into a single risk prediction of high/low risk, which can further improve HCM health management initiative [32]. To achieve this, a combination of models' output analysis and domain experts' input would be required. Finally, further ways for improvement of predictive accuracy will be tested (additional predictive models and feature selection techniques, including deep learning), as well as determining the reasons for the experts' disagreement with some of the explanation components.

### Conclusion

Although ML can have limitations in medicine [2], in this work, we showed the importance of using computer models in cardiology by predicting disease progression of HCM patients 10 years ahead, which could be used to prevent SCD. Additionally, the results confirmed findings in Chen et al [44], Gu et al [45], and Bai et al [46] that additional artificial data and semisupervised learning can provide additional low-cost and low-risk data using already available medical knowledge, increasing the predictive performance. Simple explanations of predictions contribute to the trust of provided predictions and ease the decision of experts. We hope that our work will further contribute to the goal of developing constructive strategies to prevent SCD in patients with HCM, as motivated by Maron et al [36].



## Acknowledgments

This project received funding from the European Union's Horizon 2020 research and innovation program (grant agreement no. 777204; [www.silicofcm.eu](http://www.silicofcm.eu)). This paper reflects only the authors' views. The European Commission is not responsible for any use that may be made of the information it contains.

## Conflicts of Interest

None declared.

### Multimedia Appendix 1

Bar graphs of parameter influence for each model used.

[[PNG File, 257 KB - medinform\\_v10i2e30483\\_app1.png](#)]

### Multimedia Appendix 2

A sample of the questionnaire for the first patient.

[[PDF File \(Adobe PDF File\), 597 KB - medinform\\_v10i2e30483\\_app2.pdf](#)]

## References

1. Cuocolo R, Perillo T, De Rosa E, Ugga L, Petretta M. Current applications of big data and machine learning in cardiology. *J Geriatr Cardiol* 2019 Aug;16(8):601-607 [[FREE Full text](#)] [doi: [10.11909/j.issn.1671-5411.2019.08.002](https://doi.org/10.11909/j.issn.1671-5411.2019.08.002)] [Medline: [31555327](#)]
2. Shameer K, Johnson K, Glicksberg B, Dudley J, Sengupta P. Machine learning in cardiovascular medicine: are we there yet? *Heart* 2018 Jul;104(14):1156-1164. [doi: [10.1136/heartjnl-2017-311198](https://doi.org/10.1136/heartjnl-2017-311198)] [Medline: [29352006](#)]
3. Fisher CK, Smith AM, Walsh JR, Coalition Against Major Diseases. Machine learning for comprehensive forecasting of Alzheimer's disease progression. *Sci Rep* 2019 Sep 20;9(1):13622 [[FREE Full text](#)] [doi: [10.1038/s41598-019-49656-2](https://doi.org/10.1038/s41598-019-49656-2)] [Medline: [31541187](#)]
4. Lee G, Nho K, Kang B, Sohn K, Kim D. Predicting Alzheimer's disease progression using multi-modal deep learning approach. *Sci Rep* 2019 Feb 13;9(1):1952 [[FREE Full text](#)] [doi: [10.1038/s41598-018-37769-z](https://doi.org/10.1038/s41598-018-37769-z)] [Medline: [30760848](#)]
5. Nilashi M, Ibrahim O, Ahmadi H, Shahmoradi L, Farahmand M. A hybrid intelligent system for the prediction of Parkinson's disease progression using machine learning techniques. *Biocybern Biomed Eng* 2018;38(1):1-15. [doi: [10.1016/j.bbe.2017.09.002](https://doi.org/10.1016/j.bbe.2017.09.002)]
6. Pinto MF, Oliveira H, Batista S, Cruz L, Pinto M, Correia I, et al. Prediction of disease progression and outcomes in multiple sclerosis with machine learning. *Sci Rep* 2020 Dec 03;10(1):21038 [[FREE Full text](#)] [doi: [10.1038/s41598-020-78212-6](https://doi.org/10.1038/s41598-020-78212-6)] [Medline: [33273676](#)]
7. Xiao J, Ding R, Xu X, Guan H, Feng X, Sun T, et al. Comparison and development of machine learning tools in the prediction of chronic kidney disease progression. *J Transl Med* 2019 Apr 11;17(1):119-113 [[FREE Full text](#)] [doi: [10.1186/s12967-019-1860-0](https://doi.org/10.1186/s12967-019-1860-0)] [Medline: [30971285](#)]
8. Konerman MA, Beste LA, Van T, Liu B, Zhang X, Zhu J, et al. Machine learning models to predict disease progression among veterans with hepatitis C virus. *PLoS One* 2019 Jan 4;14(1):e0208141 [[FREE Full text](#)] [doi: [10.1371/journal.pone.0208141](https://doi.org/10.1371/journal.pone.0208141)] [Medline: [30608929](#)]
9. Lyon A, Mincholé A, Bueno-Orovio A, Rodriguez B. Improving the clinical understanding of hypertrophic cardiomyopathy by combining patient data, machine learning and computer simulations: a case study. *Morphologie* 2019 Dec;103(343):169-179 [[FREE Full text](#)] [doi: [10.1016/j.morpho.2019.09.001](https://doi.org/10.1016/j.morpho.2019.09.001)] [Medline: [31570308](#)]
10. Maron BJ, Gardin JM, Flack JM, Gidding SS, Kurosaki TT, Bild DE. Prevalence of hypertrophic cardiomyopathy in a general population of young adults. Echocardiographic analysis of 4111 subjects in the CARDIA Study. Coronary artery risk development in (young) adults. *Circulation* 1995 Aug 15;92(4):785-789. [doi: [10.1161/01.cir.92.4.785](https://doi.org/10.1161/01.cir.92.4.785)] [Medline: [7641357](#)]
11. Olivotto I, Cecchi F, Poggesi C, Yacoub MH. Patterns of disease progression in hypertrophic cardiomyopathy: an individualized approach to clinical staging. *Circ Heart Fail* 2012 Jul 01;5(4):535-546. [doi: [10.1161/CIRCHEARTFAILURE.112.967026](https://doi.org/10.1161/CIRCHEARTFAILURE.112.967026)] [Medline: [22811549](#)]
12. O'Mahony C, Jichi F, Pavlou M, Monserrat L, Anastasakis A, Rapezzi C, Hypertrophic Cardiomyopathy Outcomes Investigators. A novel clinical risk prediction model for sudden cardiac death in hypertrophic cardiomyopathy (HCM risk-SCD). *Eur Heart J* 2014 Aug 07;35(30):2010-2020. [doi: [10.1093/eurheartj/eh439](https://doi.org/10.1093/eurheartj/eh439)] [Medline: [24126876](#)]
13. Andy AU, Guntuku SC, Adusumalli S, Asch DA, Groeneveld PW, Ungar LH, et al. Predicting cardiovascular risk using social media data: performance evaluation of machine-learning models. *JMIR Cardio* 2021 Feb 19;5(1):e24473 [[FREE Full text](#)] [doi: [10.2196/24473](https://doi.org/10.2196/24473)] [Medline: [33605888](#)]
14. Lundberg S, Allen P. A unified approach to interpreting model predictions. 2017 Dec Presented at: Proceedings of the 31st International Conference on Neural Information Processing Systems; Dec 4, 2017; Long Beach, CA, USA p. 4768-4777.

15. Magoulas G, Prentza A. Machine learning in medical applications. In: Lecture Notes in Computer Science. Berlin, Heidelberg: Springer; Sep 20, 2001:300-307.
16. Wang X, Sontag D, Wang F. Unsupervised learning of disease progression models. 2014 Presented at: Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining; Aug 2014; New York, NY, USA p. 85-94. [doi: [10.1145/2623330.2623754](https://doi.org/10.1145/2623330.2623754)]
17. Juarez-Orozco LE, Martinez-Manzanera O, Storti AE, Knuuti J. Machine learning in the evaluation of myocardial ischemia through nuclear cardiology. *Curr Cardiovasc Imaging Rep* 2019 Feb 9;12(2):1-8. [doi: [10.1007/s12410-019-9480-x](https://doi.org/10.1007/s12410-019-9480-x)]
18. Sardar P, Abbott JD, Kundu A, Aronow HD, Granada JF, Giri J. Impact of artificial intelligence on interventional cardiology: from decision-making aid to advanced interventional procedure assistance. *JACC Cardiovasc Interv* 2019 Jul 22;12(14):1293-1303 [FREE Full text] [doi: [10.1016/j.jcin.2019.04.048](https://doi.org/10.1016/j.jcin.2019.04.048)] [Medline: [31320024](https://pubmed.ncbi.nlm.nih.gov/31320024/)]
19. Elul Y, Rosenberg AA, Schuster A, Bronstein AM, Yaniv Y. Meeting the unmet needs of clinicians from AI systems showcased for cardiology with deep-learning-based ECG analysis. *Proc Natl Acad Sci U S A* 2021 Jun 15;118(24):e2020620118 [FREE Full text] [doi: [10.1073/pnas.2020620118](https://doi.org/10.1073/pnas.2020620118)] [Medline: [34099565](https://pubmed.ncbi.nlm.nih.gov/34099565/)]
20. Ding Y, Sohn JH, Kawczynski MG, Trivedi H, Harnish R, Jenkins NW, et al. A deep learning model to predict a diagnosis of Alzheimer disease by using F-FDG PET of the brain. *Radiology* 2019 Feb;290(2):456-464 [FREE Full text] [doi: [10.1148/radiol.2018180958](https://doi.org/10.1148/radiol.2018180958)] [Medline: [30398430](https://pubmed.ncbi.nlm.nih.gov/30398430/)]
21. Seah J, Tang J, Kitchen A, Gaillard F, Dixon A. Chest radiographs in congestive heart failure: visualizing neural network learning. *Radiology* 2019 Feb;290(2):514-522. [doi: [10.1148/radiol.2018180887](https://doi.org/10.1148/radiol.2018180887)] [Medline: [30398431](https://pubmed.ncbi.nlm.nih.gov/30398431/)]
22. Ohta Y, Yunaga H, Kitao S, Fukuda T, Ogawa T. Detection and classification of myocardial delayed enhancement patterns on MR images with deep neural networks: a feasibility study. *Radiol Artif Intell* 2019 May;1(3):e180061 [FREE Full text] [doi: [10.1148/ryai.2019180061](https://doi.org/10.1148/ryai.2019180061)] [Medline: [33937791](https://pubmed.ncbi.nlm.nih.gov/33937791/)]
23. Attia ZI, Kapa S, Lopez-Jimenez F, McKie PM, Ladewig DJ, Satam G, et al. Screening for cardiac contractile dysfunction using an artificial intelligence-enabled electrocardiogram. *Nat Med* 2019 Jan;25(1):70-74. [doi: [10.1038/s41591-018-0240-2](https://doi.org/10.1038/s41591-018-0240-2)] [Medline: [30617318](https://pubmed.ncbi.nlm.nih.gov/30617318/)]
24. Kwon J, Lee Y, Lee Y, Lee S, Park J. An algorithm based on deep learning for predicting in-hospital cardiac arrest. *JAMA* 2018 Jul 03;320(13):e008678. [doi: [10.1161/jaha.118.008678](https://doi.org/10.1161/jaha.118.008678)]
25. Adler ED, Voors AA, Klein L, Macheret F, Braun OO, Urey MA, et al. Improving risk prediction in heart failure using machine learning. *Eur J Heart Fail* 2020 Jan 12;22(1):139-147 [FREE Full text] [doi: [10.1002/ehf.1628](https://doi.org/10.1002/ehf.1628)] [Medline: [31721391](https://pubmed.ncbi.nlm.nih.gov/31721391/)]
26. Daghistani TA, Elshawi R, Sakr S, Ahmed AM, Al-Thwayee A, Al-Mallah MH. Predictors of in-hospital length of stay among cardiac patients: a machine learning approach. *Int J Cardiol* 2019 Aug 01;288:140-147. [doi: [10.1016/j.ijcard.2019.01.046](https://doi.org/10.1016/j.ijcard.2019.01.046)] [Medline: [30685103](https://pubmed.ncbi.nlm.nih.gov/30685103/)]
27. Gjoreski M, Gradišek A, Gams M, Simjanoska M, Peterlin A, Poglajen G. Chronic heart failure detection from heart sounds using a stack of machine-learning classifiers. : *IEEE*; 2017 Aug 21 Presented at: 13th International Conference on Intelligent Environments; 2017; Seoul, South Korea p. 14. [doi: [10.1109/ie.2017.19](https://doi.org/10.1109/ie.2017.19)]
28. Oguz C, Sen SK, Davis AR, Fu Y, O'Donnell CJ, Gibbons GH. Genotype-driven identification of a molecular network predictive of advanced coronary calcium in ClinSeq® and Framingham Heart Study cohorts. *BMC Syst Biol* 2017 Oct 26;11(1):99-14 [FREE Full text] [doi: [10.1186/s12918-017-0474-5](https://doi.org/10.1186/s12918-017-0474-5)] [Medline: [29073909](https://pubmed.ncbi.nlm.nih.gov/29073909/)]
29. Rowin E, Maron M. The role of cardiac MRI in the diagnosis and risk stratification of hypertrophic cardiomyopathy. *Arrhythmia Electrophysiol Rev Internet Radcliffe Cardiology* Dec 2016;1. [doi: [10.15420/aer.2016.13.3](https://doi.org/10.15420/aer.2016.13.3)]
30. Hoey ETD, Teoh JK, Das I, Ganeshan A, Simpson H, Watkin RW. The emerging role of cardiovascular MRI for risk stratification in hypertrophic cardiomyopathy. *Clin Radiol* 2014 Mar;69(3):221-230. [doi: [10.1016/j.crad.2013.11.012](https://doi.org/10.1016/j.crad.2013.11.012)] [Medline: [24412355](https://pubmed.ncbi.nlm.nih.gov/24412355/)]
31. Hiemstra YL, Debonnaire P, Bootsma M, Schalijs MJ, Bax JJ, Delgado V, et al. Prevalence and prognostic implications of right ventricular dysfunction in patients with hypertrophic cardiomyopathy. *Am J Cardiol* 2019 Aug 15;124(4):604-612. [doi: [10.1016/j.amjcard.2019.05.021](https://doi.org/10.1016/j.amjcard.2019.05.021)] [Medline: [31204037](https://pubmed.ncbi.nlm.nih.gov/31204037/)]
32. Just E. Understanding Risk Stratification, Comorbidities, and the Future of Healthcare. 2014. URL: <https://www.slideshare.net/healthcatalyst1/understanding-risk-stratification-comorbidities-and-the-future-of-healthcare> [accessed 2015-06-05]
33. Hess OM. Risk stratification in hypertrophic cardiomyopathy. *J Am Coll Cardiol* 2003 Sep;42(5):880-881. [doi: [10.1016/s0735-1097\(03\)00838-6](https://doi.org/10.1016/s0735-1097(03)00838-6)]
34. Steriotis A, Sharma S. Risk stratification in hypertrophic cardiomyopathy. *Eur Cardiol* 2015 Jul;10(1):31-36 [FREE Full text] [doi: [10.15420/ecr.2015.10.01.31](https://doi.org/10.15420/ecr.2015.10.01.31)] [Medline: [30310420](https://pubmed.ncbi.nlm.nih.gov/30310420/)]
35. Maron BJ, Rowin EJ, Casey SA, Haas TS, Chan RH, Udelson JE, et al. Risk stratification and outcome of patients with hypertrophic cardiomyopathy ≥60 years of age. *Circulation* 2013 Feb 05;127(5):585-593. [doi: [10.1161/circulationaha.112.136085](https://doi.org/10.1161/circulationaha.112.136085)]
36. Maron MS, Rowin EJ, Wessler BS, Mooney PJ, Fatima A, Patel P, et al. Enhanced American College of Cardiology/American Heart Association strategy for prevention of sudden cardiac death in high-risk patients with hypertrophic cardiomyopathy. *JAMA Cardiol* 2019 Jul 01;4(7):644-657 [FREE Full text] [doi: [10.1001/jamacardio.2019.1391](https://doi.org/10.1001/jamacardio.2019.1391)] [Medline: [31116360](https://pubmed.ncbi.nlm.nih.gov/31116360/)]
37. Gersh B, Maron B, Bonow R, Dearani J, Fifer M, Link M, American College of Cardiology Foundation/American Heart Association Task Force on Practice Guidelines, American Association for Thoracic Surgery, American Society of

- Echocardiography, American Society of Nuclear Cardiology, Heart Failure Society of America, Heart Rhythm Society, Society for Cardiovascular Angiography/Interventions, Society of Thoracic Surgeons. 2011 ACCF/AHA guideline for the diagnosis and treatment of hypertrophic cardiomyopathy: a report of the American College of Cardiology Foundation/American Heart Association Task Force on Practice Guidelines. *Circulation* 2011 Dec 13;124(24):e783-e831. [doi: [10.1161/CIR.0b013e318223e2bd](https://doi.org/10.1161/CIR.0b013e318223e2bd)] [Medline: [22068434](https://pubmed.ncbi.nlm.nih.gov/22068434/)]
38. Robnik-Šikonja M, Kononenko I. An adaptation of Relief for attribute estimation in regression. 1997 Presented at: Proceedings of the Fourteenth International Conference on Machine Learning (ICML) 1997; 1997; Nashville, TN, USA p. 296-304.
  39. Flaxman AD, Vahdatpour A, Green S, James SL, Murray CJ, Population Health Metrics Research Consortium (PHMRC). Random forests for verbal autopsy analysis: multisite validation study using clinical diagnostic gold standards. *Popul Health Metr* 2011 Aug 04;9:29-11 [FREE Full text] [doi: [10.1186/1478-7954-9-29](https://doi.org/10.1186/1478-7954-9-29)] [Medline: [21816105](https://pubmed.ncbi.nlm.nih.gov/21816105/)]
  40. Malley JD, Kruppa J, Dasgupta A, Malley KG, Ziegler A. Probability machines: consistent probability estimation using nonparametric learning machines. *Methods Inf Med* 2012;51(1):74-81 [FREE Full text] [doi: [10.3414/ME00-01-0052](https://doi.org/10.3414/ME00-01-0052)] [Medline: [21915433](https://pubmed.ncbi.nlm.nih.gov/21915433/)]
  41. Abraham A, Pedregosa F, Eickenberg M, Gervais P, Mueller A, Kossaifi J, et al. Machine learning for neuroimaging with scikit-learn. *Front Neuroinform* 2014 Feb 21;8:14 [FREE Full text] [doi: [10.3389/fninf.2014.00014](https://doi.org/10.3389/fninf.2014.00014)] [Medline: [24600388](https://pubmed.ncbi.nlm.nih.gov/24600388/)]
  42. Chen T, Guestrin C. XGBoost: a scalable tree boosting system. USA: Association for Computing Machinery; 2016 Presented at: Proceedings of the 22th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining KDD '16; Aug 13, 2016; San Francisco, CA, USA p. 785-794. [doi: [10.1145/2939672.2939785](https://doi.org/10.1145/2939672.2939785)]
  43. O'Malley T, Bursztein E, Long J, Chollet F, Jin H, Invernizzi L. keras-team/keras-tuner: Hyperparameter Tuning for Humans. URL: <https://github.com/keras-team/keras-tuner> [accessed 2021-03-05]
  44. Chen S, Bortsova G, Juárez A, van TG, de BM. Multi-task attention-based semi-supervised learning for medical image segmentation. *Lect Notes Comput Sci* 2019:457-465. [doi: [10.1007/978-3-030-32248-9\\_51](https://doi.org/10.1007/978-3-030-32248-9_51)]
  45. Gu L, Zhang X, You S, Zhao S, Liu Z, Harada T. Semi-supervised learning in medical images through graph-embedded random forest. *Front Neuroinform* 2020 Nov 10;14:601829 [FREE Full text] [doi: [10.3389/fninf.2020.601829](https://doi.org/10.3389/fninf.2020.601829)] [Medline: [33240071](https://pubmed.ncbi.nlm.nih.gov/33240071/)]
  46. Bai W, Oktay O, Sinclair M, Suzuki H, Rajchl M, Tarroni G. Semi-supervised learning for network-based cardiac MR image segmentation. In: *Lecture Notes in Computer Science*. Cham: Springer; 2017:253-260.
  47. Viceconti M, Henney A, Morley-Fletcher E. In silico clinical trials: how computer simulation will transform the biomedical industry. *Int J Clin Trials* 2016 May 09;3(2):37. [doi: [10.18203/2349-3259.ijct20161408](https://doi.org/10.18203/2349-3259.ijct20161408)]
  48. Niederer SA, Lumens J, Trayanova NA. Computational models in cardiology. *Nat Rev Cardiol* 2019 Feb;16(2):100-111 [FREE Full text] [doi: [10.1038/s41569-018-0104-y](https://doi.org/10.1038/s41569-018-0104-y)] [Medline: [30361497](https://pubmed.ncbi.nlm.nih.gov/30361497/)]
  49. Tannenbaum SJ, Holford NHG, Lee H, Peck CC, Mould DR. Simulation of correlated continuous and categorical variables using a single multivariate distribution. *J Pharmacokinet Pharmacodyn* 2006 Dec;33(6):773-794. [doi: [10.1007/s10928-006-9033-1](https://doi.org/10.1007/s10928-006-9033-1)] [Medline: [17053984](https://pubmed.ncbi.nlm.nih.gov/17053984/)]
  50. Robnik Šikonja M. Dataset comparison workflows. *IJDS* 2018;3(2):126-145. [doi: [10.1504/ijds.2018.10013385](https://doi.org/10.1504/ijds.2018.10013385)]
  51. Bøtcher SG, Dethlefsen C. DEAL: a package for learning Bayesian networks. In: *J Stat Softw*. Aalborg University: Department of Mathematical Sciences; 2003.
  52. D'Agostino R, Stephens M. Goodness-of-Fit Techniques. New York, NY, USA: Marcel Dekker; 1986.
  53. Robnik-Šikonja M, Bohanec M. Perturbation-Based Explanations of Prediction Models. Cham: Springer; 2018:159-175.
  54. SHaRe. URL: <https://theshareregistry.org/home/patient> [accessed 2021-05-14]

## Abbreviations

- AI:** artificial intelligence
- BSA:** body surface area
- CMR:** cardiac magnetic resonance
- COPD:** chronic obstructive pulmonary disease
- ECG:** electrocardiogram
- Echo:** echocardiogram
- GB:** gradient boosted
- HCM:** hypertrophic cardiomyopathy
- KNN:** k-nearest neighbors
- LA\_d:** left atrial diameter
- LA\_Vol:** left atrial volume
- LR:** linear regression
- LVEF:** left ventricular ejection fraction LVIDd: left ventricular internal diameter at end diastole
- LVIDs:** left ventricular internal diameter at end systole
- MAE:** mean absolute error
- ML:** machine learning

**MVND:** multivariate normal distribution  
**NN:** neural network  
**NYHA:** New York Heart Association  
**RF:** random forest  
**RMSE:** root-mean-square error  
**RRMSE:** relative root-mean-square error  
**SCD:** sudden cardiac death  
**SHAP:** Shapley additive explanation  
**SVM:** support vector machine

*Edited by C Lovis; submitted 18.05.21; peer-reviewed by T Kahlon, P Li, P Bhadra; comments to author 27.09.21; revised version received 27.10.21; accepted 04.12.21; published 02.02.22.*

*Please cite as:*

*Pičulin M, Smole T, Žunkovič B, Kokalj E, Robnik-Šikonja M, Kukar M, Fotiadis DI, Pezoulas VC, Tachos NS, Barlocco F, Mazzarotto F, Popović D, Maier LS, Velicki L, Olivotto I, MacGowan GA, Jakovljević DG, Filipović N, Bosnić Z*

*Disease Progression of Hypertrophic Cardiomyopathy: Modeling Using Machine Learning*

*JMIR Med Inform 2022;10(2):e30483*

URL: <https://medinform.jmir.org/2022/2/e30483>

doi: [10.2196/30483](https://doi.org/10.2196/30483)

PMID: [35107432](https://pubmed.ncbi.nlm.nih.gov/35107432/)

©Matej Pičulin, Tim Smole, Bojan Žunkovič, Enja Kokalj, Marko Robnik-Šikonja, Matjaž Kukar, Dimitrios I Fotiadis, Vasileios C Pezoulas, Nikolaos S Tachos, Fausto Barlocco, Francesco Mazzarotto, Dejana Popović, Lars S Maier, Lazar Velicki, Iacopo Olivotto, Guy A MacGowan, Djordje G Jakovljević, Nenad Filipović, Zoran Bosnić. Originally published in JMIR Medical Informatics (<https://medinform.jmir.org>), 02.02.2022. This is an open-access article distributed under the terms of the Creative Commons Attribution License (<https://creativecommons.org/licenses/by/4.0/>), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Medical Informatics, is properly cited. The complete bibliographic information, a link to the original publication on <https://medinform.jmir.org/>, as well as this copyright and license information must be included.

Original Paper

# Early Identification of Maternal Cardiovascular Risk Through Sourcing and Preparing Electronic Health Record Data: Machine Learning Study

Nawar Shara<sup>1</sup>, PhD; Kelley M Anderson<sup>2</sup>, PhD, FNP; Noor Falah<sup>3</sup>, MS; Maryam F Ahmad<sup>3</sup>; Darya Tavazoei<sup>3</sup>; Justin M Hughes<sup>3</sup>, MS; Bethany Talmadge<sup>4</sup>, BS; Samantha Crovatt<sup>4</sup>, BS; Ramon Dempers<sup>4</sup>, MSc

<sup>1</sup>MedStar Health Research Institute, Georgetown-Howard Universities Center for Clinical and Translational Science, Hyattsville, MD, United States

<sup>2</sup>Georgetown University, Washington, DC, United States

<sup>3</sup>MedStar Health Research Institute, Hyattsville, MD, United States

<sup>4</sup>Invaryant Inc, Roswell, GA, United States

**Corresponding Author:**

Nawar Shara, PhD

MedStar Health Research Institute

Georgetown-Howard Universities Center for Clinical and Translational Science

6525 Belcrest Road

Hyattsville, MD, 20782

United States

Phone: 1 240 618 7859

Email: [Nawar.Shara@Medstar.net](mailto:Nawar.Shara@Medstar.net)

## Abstract

**Background:** Health care data are fragmenting as patients seek care from diverse sources. Consequently, patient care is negatively impacted by disparate health records. Machine learning (ML) offers a disruptive force in its ability to inform and improve patient care and outcomes. However, the differences that exist in each individual's health records, combined with the lack of health data standards, in addition to systemic issues that render the data unreliable and that fail to create a single view of each patient, create challenges for ML. Although these problems exist throughout health care, they are especially prevalent within maternal health and exacerbate the maternal morbidity and mortality crisis in the United States.

**Objective:** This study aims to demonstrate that patient records extracted from the electronic health records (EHRs) of a large tertiary health care system can be made actionable for the goal of effectively using ML to identify maternal cardiovascular risk before evidence of diagnosis or intervention within the patient's record. Maternal patient records were extracted from the EHRs of a large tertiary health care system and made into patient-specific, complete data sets through a systematic method.

**Methods:** We outline the effort that was required to define the specifications of the computational systems, the data set, and access to relevant systems, while ensuring that data security, privacy laws, and policies were met. Data acquisition included the concatenation, anonymization, and normalization of health data across multiple EHRs in preparation for their use by a proprietary risk stratification algorithm designed to establish patient-specific baselines to identify and establish cardiovascular risk based on deviations from the patient's baselines to inform early interventions.

**Results:** Patient records can be made actionable for the goal of effectively using ML, specifically to identify cardiovascular risk in pregnant patients.

**Conclusions:** Upon acquiring data, including their concatenation, anonymization, and normalization across multiple EHRs, the use of an ML-based tool can provide early identification of cardiovascular risk in pregnant patients.

(*JMIR Med Inform* 2022;10(2):e34932) doi:[10.2196/34932](https://doi.org/10.2196/34932)

**KEYWORDS**

electronic health record; maternal health; machine learning; maternal morbidity and mortality; cardiovascular risk; data transformation; extract; transform; load; artificial intelligence; electronic medical record

## Introduction

### Background

Each year in the United States, maternal morbidity and mortality (MMM) accounts for more than 700 deaths and an additional 50,000 life-threatening complications associated with pregnancy and childbirth [1]. It is estimated that 70% of these events are preventable [1]. Cardiovascular disease accounts for 60% of maternal morbidity events and over one-third of maternal deaths in the United States [2]. More than 50% of MMM events are attributed to cardiovascular causes including cardiomyopathy (11.5%), thrombotic pulmonary embolism (9.6%), cerebrovascular accidents (8.2%), hypertensive disorders of pregnancy (6.6%), and other cardiovascular conditions (15.5%) [1]. In 2019, the national maternal death rate was 20.1 deaths per 100,000 live births [3]. It is estimated that 68.2% of pregnancy-related cardiovascular deaths are preventable [4].

Exacerbating the crisis, health care data are fragmented as patients seek care from diverse sources, including different health care systems and telehealth providers. Consequently, coordinating patients' care with disparate health records continues to increase in complexity.

We hypothesize that a systematic method for identifying risk early by analyzing changes in a patient's health data based on complete data set trends is possible and facilitates early intervention and treatment of high-risk conditions in pregnant women. Early identification and intervention of these conditions would likely result in a measurable reduction in maternal fatalities and life-threatening complications.

A previous study focused on predicting common maternal postpartum complications by leveraging machine learning (ML) and electronic health records (EHRs) highlighted the risk level of maternal postpartum complications requiring inpatient care [5]. Data were gathered from patients' dates of gestation to delivery and demonstrated that routinely collected health data, when used in conjunction with ML, have the potential to accurately predict postpartum outcomes [5].

With this as our basis, our aim is to demonstrate that ML and aggregated EHRs can be leveraged to surface signals and trends in patients' medical records to identify predictors of cardiovascular conditions during pregnancy.

Through a retrospective study based on a cohort from a large tertiary health care system of 32,409 patients who were seen during pregnancy, we demonstrate that Invariant's Health Outcomes for all Pregnancy Experiences–Cardiovascular-Risk Assessment Technology (HOPE-CAT), an ML-based risk assessment algorithm, identifies factors that may indicate the development of cardiovascular conditions that lead to MMM.

### Overview—Data and ML

ML is becoming a disruptive force in health care, and its application is broad, including imaging, risk identification, and risk assessment to inform and improve patient care and outcomes [6]. Recent studies have demonstrated that ML, compared with traditional statistical modeling, is a more effective tool in predicting sex-specific and cardiovascular diseases [7]. In

addition, when combined with traditional logical regression, ML may assist in identifying novel predictors of disease [8].

However, the vast differences that exist in each individual's corpus of health data and the lack of standards to define the capture of data create challenges for ML. In addition to complexity and variation among patients, there are systemic issues that render the data unreliable and fail to create a single accurate view of each patient. Despite standardization efforts, including Fast Healthcare Interoperability Resources and Continuation of Care Documents, adoption of these standards and upgrades is slow. In addition, within health care, patients often receive care from different providers and specialists for various conditions, obtaining diverse medications and treatments without a *clearinghouse* to ensure that all providers have access to all relevant data. Finally, owing to lack of standardization, when data are sourced from disparate systems, the resulting data must be cleaned and normalized to be made actionable. The lack of connectivity in health care creates challenges for providers, who provide care with limited and often incomplete patient information.

It is important to note that for this study, all available data were sourced from a single health care system with 10 hospitals in addition to outpatient clinics, which was both an advantage and a challenge. One advantage was that the data set represented a diverse patient population in a system with many hospitals and outpatient clinics. However, because there was only access to the single system, if a patient sought care at an external facility, the data from those visits were not available in the data set. This demonstrates one of the major challenges of the aforementioned lack of interconnectivity within the US health care system.

Despite the defined systemic problems, ML models have several advantages for the assimilation and evaluation of complex health care data. Unlike traditional statistical models, ML offers flexibility and scalability, which makes it deployable for many tasks, such as risk stratification, diagnosis and classification, and survival intervention [9]. However, when considering the use of these tools for health care data, one must understand that there are limitations to be anticipated and considered. Ethically speaking, notwithstanding the systemic issues described, clinical implementation of the technology must be for the direct benefit of a patient and their providers. The completeness of data cannot be assured, nor can it be assumed that those data are always accurately captured; additionally, the ethical use of these technologies mandates respect for patients' sensitive personal health information throughout their use.

## Methods

### Technology

For this project, the following software and platforms were used: 4 Cerner Millennium (edition 2018.01) electronic medical record software, PeriBirth (PeriGen), R (version 4.05; R Foundation for Statistical Computing); Microsoft Azure Cloud, Microsoft Azure Data Studio, Microsoft Azure Machine Learning Studio, virtual machine, Microsoft SQL Server Management Studio, Invariant's health platform, and HOPE-CAT.

For the purpose of this study, HOPE-CAT analyzes an individual patient's EHR data on an encounter-by-encounter basis to identify risk factors (eg, elevated blood pressure readings, shortness of breath, and chest pain) indicative of the development or worsening of cardiovascular conditions.

HOPE-CAT was trained via causal inference, with limited training supervision, using established maternal cardiovascular risk factors and covariates, such as physical findings, symptoms, and medical history ([Textbox 1](#)).

**Textbox 1.** Risk factors and covariates used to train the Health Outcomes for all Pregnancy Experiences–Cardiovascular-Risk Assessment Technology.

- **Symptoms (variable risks)**
  - Dyspnea (red flag risk)
  - Orthopnea (red flag risk)
  - Tachypnea
  - Asthma unresponsive to therapy
  - Swelling in face or hands
  - New or worsening headache
  - Heart palpitations
  - Dizziness or syncope
  - Chest pain
- **Physical findings (variable risks)**
  - Loud heart murmur
  - Basilar crackles in lungs
  - Resting heart rate  $\geq 120$  beats per minute (red flag risk)  $\geq 110$  beats per minute
  - Systolic blood pressure  $\geq 160$  mm Hg (red flag risk)  $\geq 140$  mm Hg
  - Respiratory rate  $\geq 30$  (red flag risk)  $\geq 24$
  - Oxygen saturation  $\leq 94\%$  (red flag risk)  $\leq 96\%$
- **Medical history (static risks)**
  - Aged  $\geq 40$  years
  - Race=African American
  - Prepregnancy obesity (BMI  $\geq 35$ )
  - Prepregnancy diagnosis of diabetes
  - Prepregnancy diagnosis of hypertension
  - Substance use (nicotine, cocaine, alcohol, and methamphetamines)
  - History of chemotherapy
  - History of complications in labor or delivery
  - History of heart disease

HOPE-CAT was then used to simulate chronological patient encounters as they occurred in the medical records. The onsets of risk detected by HOPE-CAT were compared with EHR-recorded diagnoses or interventions in the source data's timeline. Loss vectoring methods were used to determine the delta, or difference, between HOPE-CAT's outputs and the anticipated outputs, thereby guiding the learning and training. In this study, the patient encounters and outcomes were already known, and HOPE-CAT was configured to simulate patient encounters (eg, clinic appointments, emergency department visits, and hospital admissions) on the encounter dates recorded in each patient's EHR to assess the available data and detect potential risks. A delta was then determined between

HOPE-CAT's assessments and actions taken by the health care provider on the same dates with the same information.

### Criteria and Requirements Assessment

Inclusion and exclusion criteria to be pulled from the EHR were defined to create the data set to ensure that the algorithms had adequate data to analyze for trends and were able to designate risk profiles as early in the process as possible for each patient. Inclusion and exclusion criteria were agreed upon by clinical cardiovascular and maternal health experts and data scientists. These inclusion criteria included patient demographics (eg, age, race, and geographic location), physiologic measures (eg, blood pressure, heart rate, and oxygen saturation), symptoms (eg,

headache and shortness of breath), and health history from each patient encounter during pregnancy. Once criteria were defined, the list of variables were organized to identify the sources of data required.

## Data Sourcing, Cleansing, Scrubbing, and Normalization

### Data Acquisition (Institutional Policy Compliance)

For this study, institutional privacy and security policies were followed to ensure that patient data were protected and secure throughout the project. Institutional review board approval was obtained from the study institution. Analysts handling the data maintain standing access to various databases containing patient-EHR and other research data. Access is individualized and maintained through the institution's active directory. All activities within these systems are tracked and auditable, and institutional review board approval is required before any research-related data extraction. Data access methods, as well as data extraction, transfer, and anonymization procedures, were reviewed by the institution's data security team before the creation of the shared analysis environment to ensure all necessary security requirements were implemented before the release of data.

### Data Sources and Extraction

There were multiple data sources within the hospital system, each with its own access restrictions, and in some cases, data sources were administered by different departments or groups within the hospital. The 2 primary systems used were a direct connection to Cerner's underlying Oracle database, as well as an enterprise data warehouse (EDW) solution, which contains data from the Cerner EHR, as well as other third-party billing, quality, and safety systems employed at the different member institutions. Access to both systems was controlled through specific roles defined in the active directory, the Microsoft Lightweight Directory Access Protocol service. RStudio (running R version 4.05) was used to query both data sources for extraction, as well as for subsequent data transformation. It should be noted that it is theoretically possible to perform discovery and extraction of this data using Cerner's supplied suite of tools (eg, Discern Analytics and Cerner Command Language); however, because of the large number of variables and size of the data set, having other solutions available provided a significant advantage, both in performance and ease of use.

Both EDW and Cerner use a relational database architecture. Cerner's data model is primarily *visit-centric*, which means that most data created within the EHR tie together via a unique encounter ID that is created for each visit. Visits connect together through a unique person ID, and certain tables—such as the address table, family history table, and problems table (for chronic conditions)—are kept at the *person* grain. The EDW keeps these source identifiers and also includes additional fields to allow for cross-walking of visits and patients between the different imported data sources.

The data selection process began with a baseline population of patients who had a documented delivery between January 1, 2017, and December 31, 2020. A delivery was defined as a documented delivery procedure as outlined by the Centers for

Disease Control and Prevention [1]. From the initial population, International Classification of Diseases, 10th Revision (ICD-10), diagnosis codes were used to identify those patients' prenatal visits. After compiling the initial list of visits for each patient, diagnoses, selected clinical variables, and personal information (demographics) were abstracted for each patient.

In addition, visits created because of a historical upload or import from another source were excluded. These visits had registration dates starting in 1900. An age filter was also implemented so that only data from patients aged 18-35 years at the time of the encounter were received. *Visit* entries that were created because of communication between staff and patients, such as patient portal messages or phone calls, were removed if there was no relevant clinical data or if the data otherwise did not meet the established inclusion criteria.

The EDW was used for supplemental data not housed within the main EHR environment, such as diagnosis-related group codes to categorize diagnoses and complications, as well as *cleaned* versions (with duplicates removed) of certain types of data, such as medication administrations, to prevent duplicate work. A large portion of the clinical data needed, such as laboratory results, measurements, and other discrete clinical observations, were sourced from the *clinical events* table within Cerner and further categorized after extraction. The clinical events table uses the same field for result values regardless of the variable, so additional fields, such as the result *unit* (eg, lb, kg, and mm Hg), were included for additional context; this also allowed for the comparison of variables, such as weight, that can be entered as either pounds or kilograms.

This time-consuming exercise was simplified by the creation and maintenance of a comprehensive, well-documented data catalog using the tools provided in the database administration studio, which was updated when data were added from the source systems (ie, the metadata repository or data dictionary). Cerner does provide a table that has some preconfigured event categorizations in a hierarchy. However, the hierarchies and category labels are customizable at each institution, so manual review was still required to create comprehensive groups.

Validation checks were completed by manually combing through the events list to check if any code or piece of information had been missed. This step was crucial for maternal history and delivery information, as these can be documented in different ways because of various workflows across different hospitals or departments. Free-text clinical notes were not used for this study because of the additional time and computing resources that would have been required for proper removal of protected health information and identification of clinically relevant text.

A series of checks were also completed to ensure the accuracy of the data. After a variable list was developed, the data were once again validated to ensure that the variables were accounted for within the events. Randomized individual visits were then selected to check for events that would be relevant and were not already in a categorized set.

### Data Anonymization

Although the data were administratively and medically permitted to be viewed, for privacy, they were deidentified, and therefore,



variables were transformed in stages before being transferred into the dedicated Azure environment for analysis. The first set of variables transformed were basic demographic information, such as patient name and address. Zip codes were compared with census data and discarded after characterization as urban, rural, or suburban. Date time fields, such as the registration date time for visits and date time stamps associated with visits, were split into 2 fields containing the date and time of the event, respectively. The date of the index visit was calculated for each unique patient, and the date field was transformed to represent the number of calendar days from that index visit date. Time fields were kept for sequencing of events within a patient’s course. Within the system, the unique person and encounter IDs were hashed to create new person and visit IDs to prevent reidentification. A master key was created to tie patients and visits together and was only available locally to the data analysts at the home institution. The key could also be used to backtrack and revalidate in the case of errors if something did not make logical sense or if additional variables needed to be pulled after the original extraction.

The last piece of deidentification involved the events themselves. This involved parsing procedural histories and removing event types, such as comments, dates, and other free-text entry fields where identifying patient information could potentially be entered.

### Data Transformation

After completion of the initial extraction and anonymization of the data, a new database schema was necessary to house and

store the results for subsequent analysis. As a result, additional data transformation was necessary to combine different but related data elements into a single table and to aid HOPE-CAT. As stated previously, a large portion of the clinical data were further categorized after extraction. Three new fields were created to accomplish this, based on the categorization of the data fields: CATEGORY, SUB\_CATEGORY, and CLINICAL\_CAT. The field CATEGORY is the parent hierarchy, consists of values such as *event* and *diagnosis*, and signifies from which set of tables the event came. SUB\_CATEGORY is the next level down and changes based on context. In the diagnosis table, example values include *admitting diagnosis* and *discharge diagnosis*, whereas examples from the *events* table include *labs*, *measurements*, and *medication administration*. CLINICAL\_CAT is not used in all tables but provides additional categories, such as *blood gases*, *metabolic panel*, *infectious disease*, and *hematology* for lab events and *vitals* and *weight and BMI* for measurements. These 3 fields helped standardize data in different tables for easier processing.

The final database schema used a relational structure similar to that of the original tables in the EHR. All of the final tables, with the exception of the FamilyHX and Race tables, contained both the transformed person IDs and the transformed visit IDs to allow for easier analysis at either the visit or patient level. The final database schema is shown in Figure 1.

Figure 1. Database schema.



## Normalizing and Loading

Data were provided to Invariant's team in the form of a Microsoft SQL Server database, hosted on a Microsoft Azure Virtual Machine. A data catalog was created to provide the team with an inventory of available data. In addition to data profiling, statistics, and other contents, the data catalog also provided a

descriptive index pointing to the location and type of available data. Owing to the large volume of data, the tables were individually loaded to the database, using Azure Data Studio, as CSV files (Table 1). Indexes were later added to the tables on key fields to make queries necessary for analysis optimization.

**Table 1.** Database overview.

Description	Value, n (%)
Total patients in the database	32,409 (100)
Patients with at least one risk identified	18,095 (55.83)
Patients who delivered on the first visit <sup>a</sup>	14,855 (45.83)
Patients who only had 1 visit	11,485 (35.44)
Patients with <i>red flag</i> risk levels identified <sup>b</sup>	1716 (5.29)
<b>Number of births</b>	
Total births	37,457 (100)
Single live births	36,564 (97.62)
Twin births	545 (1.45)
Triplet births	13 (0.03)
Stillbirths	294 (0.78)
<b>Number of patients in top detected conditions</b>	
Preeclampsia	3468 (10.7)
Eclampsia	29 (0.09)
Cardiomyopathy	34 (0.1)
Cerebral infarction (stroke)	5 (0.02)
<b>Number of patients with static risks based on category</b>	
BMI $\geq$ 35	2800 (8.64)
African American	8194 (25.28)
History of substance use	3469 (10.7)

<sup>a</sup>These were excluded as there was no supporting retrospective data.

<sup>b</sup>Specific severe risk factors or 4 or more total risk factors.

## ML Training and Execution

### Training Networks and Building Layers

To validate HOPE-CAT against retrospective patient records through simulated patient encounters (ie, office visits) from the data, training was first completed. HOPE-CAT was trained to assess the available data chronologically by visit, as providers would have recorded them in real time. To account for the anonymization of patient-encounter dates, HOPE-CAT was trained to work using a duration function (day count), rather than a date function, to accurately determine the delta. Data collected at, and related to, each visit (eg, patient demographics, physical findings, symptoms, and medical history) were provided as input to HOPE-CAT for analysis to detect changes and trends in the patient's data. If HOPE-CAT detected risk based on the visit data and the risk factors in which it was trained (Textbox 1), a risk profile was generated for that specific patient encounter. Two types of risk profiles were generated indicating standard risk or high risk, noted as a *red flag*. Red flag risks

indicated that the patient was experiencing either single severe physiological symptoms (eg, elevated blood pressure or orthopnea) or multiple risk factors (4 or more) that may be predictors of needing immediate evaluation. A risk profile establishes that risk factors indicative of the development of severe or worsening cardiovascular conditions are present. These conditions include, but are not limited to, preeclampsia, eclampsia, peripartum cardiomyopathy, cerebral infarction, myocardial infarction, heart failure, and pulmonary embolism.

Typically, HOPE-CAT evaluates for a patient's individual baseline metrics before further analysis. For example, if a patient's systolic blood pressure baseline is lower than the medical mean, a high reading would be below the medically recommended high-risk value in cases such as preeclampsia. However, owing to the nature of retrospective data and, in many cases, the lack of medical history, establishing personal baselines for each patient was disabled for this study. Therefore, any patient whose data had fallen outside the medically accepted averages (norms) was flagged and not used for training.

Refining the data ingested during data preparation allowed for the isolation of data that related directly to patients who exhibited static risk (ie, patient information that does not change, such as race and prepregnancy history) or variable risk factors (eg, physiologic measures and symptoms). The networks were trained to identify both an individual's static risk factors and any additional variable risk factors developed over time. Throughout the training process, Periodic testing was performed. For missing data (eg, weight), those data were requested from the data sources and added to the study database, and the catalog and data dictionaries were updated. Once each issue was resolved, the ingestion and refining process was continued, and testing was repeated before additional training.

The training results were reviewed by clinical experts, and some adjustments were made in the context of *static* risk. The system was retrained to accommodate these changes, and once again, a series of manual tests were run to ensure that the changes had the appropriate effect.

As data were layered into the HOPE-CAT, outliers or patients with data that did not meet the evaluation criteria (eg, a single visit encounter was available, meaning trends could not be identified) were identified and flagged for exclusion (Table 1). It was found that certain data layers initially included in the requirements had limited use and that some of the data were held in other tables and, in some cases, in other databases.

### Testing ML Outputs

Test data based on the specified inclusion criteria from cardiovascular and maternal health experts, and findings of previous studies, were used to train HOPE-CAT using human reviews of maternal data. These metrics and parameters were loaded and run against the test data. The outputs from the test data set were reviewed manually on a patient-by-patient basis. The advantage this study had in the context of medical care is that the retrospective data had clearly defined outcomes for all the patients included in the result set, thereby allowing precise analysis of HOPE-CAT's outputs, with direct confirmations of the correlation of the defined risk to the outcome of the pregnancy. For a risk coded or identified by HOPE-CAT, it was possible to determine the accuracy of the assessment against hard data (eg, the patient being diagnosed with a cardiovascular condition).

### Running ML and Reviewing Results

Once training was complete, HOPE-CAT was run against the full data set to determine the risk level against the encounter-duration function. When HOPE-CAT identified a certain level of risk, the encounter date associated with the output was compared with the date of when a diagnosis was made or the provider intervened (ie, the delta). The delta between the detection by HOPE-CAT and the diagnosis or intervention by the provider was assessed and quantified. In most cases, HOPE-CAT had the advantage over the provider as HOPE-CAT had a single, condensed view to the patient's historical data, data trends, and micro and macro changes in the patient's health. As described earlier, the advantage of the retrospective data allowed for in-depth manual reviews of the data. The process involved reviewing each method by retrieving

the relevant data against the results of HOPE-CAT, and each result was cross-checked and tabulated. The tabulated information was then cross-checked by the independent quality team. An important part of ML is the classification of outputs, which identifies errors or artifacts that the system cannot explain. These data were flagged for human review and classification; as the system was designed to detect primarily cardiac-related events, it did not know how to classify certain events; therefore, HOPE-CAT flagged them as errors. Once reviewed by the data analysts, a set of these errors were identified as *organ failures*, and in review of the data, all references to organ failure in the data were detected, and the classification was added to the classification system. This resulted in the expansion of the classification algorithm to alert providers of the additional risk of potential organ failure in a patient, indicating that a patient may require further monitoring and intervention to prevent advancement of disease state and more severe outcomes. For example, patients with HELLP (hemolysis, elevated liver enzymes, and low platelets) syndrome or preeclampsia should be monitored for hematologic changes or changes in liver or kidney function, respectively, which may indicate disease advancement and potential organ failure. This process demonstrates that error handling is an effective tool for identifying and correcting omissions or unexpected events in the data.

## Results

This study has shown that patient records from EHRs, when aggregated, can be made actionable for the goal of effectively using ML, specifically to identify cardiovascular risk in pregnant patients. The resulting delta informs future studies in which HOPE-CAT will be deployed to monitor for and alert providers to real time trends in patient data.

## Discussion

### Limitations

Several methods used within this study are proprietary to Invariant. These methods are related to HOPE-CAT ML, the risk stratification algorithm designed to establish patient-specific baselines to identify and establish cardiovascular risk based on deviations from the patient's baseline. That said, these processes being proprietary to Invariant do not limit future research in this purview. ML processes similar to HOPE-CAT may be developed; however, the processes of training, variable weighting, and validating may differ.

### Conclusions

Within this study, 32,409 anonymized health records were extracted from multiple Cerner EHR systems. Data were collected and applied in four distinct steps: design, discovery, ingestion, and refinement. Extensive measures were taken to meet patient privacy requirements and the home institution's security requirements, including removing key identifiable data points, including names, addresses, dates of birth, and zip codes, as well as other measures to protect patient privacy. Further security measures were taken to provide access to the data and establish the environment in the Microsoft Azure Cloud while

maintaining the home institution's security policies and practices. Data were then cleaned, scrubbed, validated, structured, optimized, and normalized before setting up analytical processing capabilities. To prepare for analytical processing, iterative, layered training of samples of the data was executed, and reviewed training for the learning engine was run, to ensure an abundance of data categories were available in large enough quantities to guarantee that results were reproducible and scalable for complete analysis in a *real-world* live setting. The latter part will be vital in instances where these processes are used in vivo.

Future studies involving the HOPE-CAT may include the following: the addition of geographic data and other data related to social determinants of health, including unstructured sources (eg, chart notes, family histories, and imaging) with natural language processing or prospective in vivo application.

### Recommendations From This Study and for Future Studies

During the design process, it is recommended to consider the following:

- Are enough data available to represent the pattern of interest?
- Are the data available accurate? (Plausibility checks for accuracy, misspellings, parsing, and standardization to specific locales are recommended.)
- Have data correction schemes been considered?
- Data-cleaning decisions go beyond technical feasibility; evaluating ethical and legal implications is also necessary.
- Including an iterative review process with clinicians for algorithm-inclusion requirements is recommended.

- Educating end users on the many implications (medical, legal, and ethical) of using these technologies to inform better health outcomes and setting expectations for artificial intelligence and ML strengths and limitations are recommended.
- Training end users on how best to use artificial intelligence and ML tools and interpret outputs is recommended.

Before gathering data, source selection is key. It is important to first determine if data will need to be gathered from multiple sources, and if so, how to integrate them. Assessing the number of events required per observation period and determining beforehand how much data are needed to represent segment variability or to simply come to a successful conclusion could be very useful. Formulating an easy method for matching data from alternate sources is key to ensuring sufficient data for any project.

Data preparation methods should include formal processes, such as the creation of dictionaries, catalogs, and other controls, that allow the process to be repeatable and scalable. Metadata, persistent managed storage, and reusable transformation or cleansing, and the information around them, must be included to make data preparation efficient and consistent. Assessing how the data need to be aligned for the analysis often involves cardinality, binning, correlations, derivations of new values, gender or identity analyses, and other methods to prepare data at the needed level of granularity. Within this study, it was found that once data refinement was started, additional data were needed to better suit the purpose of the study. During the process of refining data, it is recommended that one determines how fit the data are for the intended purpose and if further data may be needed.

### Conflicts of Interest

None declared.

### References

1. Reproductive health: maternal mortality. Centers for Disease Control and Prevention. 2020. URL: <https://www.cdc.gov/reproductivehealth/maternal-mortality/index.html> [accessed 2021-08-23]
2. Petersen EE, Davis NL, Goodman D, Cox S, Mayes N, Johnston E, et al. Vital signs: pregnancy-related deaths, United States, 2011-2015, and strategies for prevention, 13 states, 2013-2017. *MMWR Morb Mortal Wkly Rep* 2019;68(18):423-429 [FREE Full text] [doi: [10.15585/mmwr.mm6818e1](https://doi.org/10.15585/mmwr.mm6818e1)] [Medline: [31071074](https://pubmed.ncbi.nlm.nih.gov/31071074/)]
3. Hoyert DL. Maternal mortality rates in the United States, 2019. Centers for disease control and prevention. Hyattsville: National Center for Health Statistics Health E-Stats; 2021. URL: <https://stacks.cdc.gov/view/cdc/103855> [accessed 2022-02-02]
4. Report from nine maternal mortality review committees. Building U.S. Capacity to Review and Prevent Maternal Deaths.: CDC Foundation; 2018. URL: <https://www.cdcfoundation.org/sites/default/files/files/ReportfromNineMMRCs.pdf> [accessed 2021-08-23]
5. Betts KS, Kisely S, Alati R. Predicting common maternal postpartum complications: leveraging health administrative data and machine learning. *BJOG* 2019;126(6):702-709. [doi: [10.1111/1471-0528.15607](https://doi.org/10.1111/1471-0528.15607)] [Medline: [30628159](https://pubmed.ncbi.nlm.nih.gov/30628159/)]
6. Callahan A, Shah NH. Machine learning in healthcare. In: Sheikh A, Bates D, Wright A, Cresswell K, editors. *Key advances in clinical informatics: transforming health care through health information technology*. Cambridge: Academic Press; 2017:279-291.
7. Sarajlic P, Plunde O, Franco-Cereceda A, Bäck M. Artificial intelligence models reveal sex-specific gene expression in aortic valve calcification. *JACC Basic Transl Sci* 2021;6(5):403-412 [FREE Full text] [doi: [10.1016/j.jacbts.2021.02.005](https://doi.org/10.1016/j.jacbts.2021.02.005)] [Medline: [34095631](https://pubmed.ncbi.nlm.nih.gov/34095631/)]

8. Sarajlic P, Simonsson M, Jernberg T, Bäck M, Hofmann R. Incidence, associated outcomes, and predictors of upper gastrointestinal bleeding following acute myocardial infarction: a SWEDHEART-based nationwide cohort study. *Eur Heart J Cardiovasc Pharmacother* (forthcoming) 2021:pwab059. [doi: [10.1093/ehjcvp/pwab059](https://doi.org/10.1093/ehjcvp/pwab059)] [Medline: [34423350](https://pubmed.ncbi.nlm.nih.gov/34423350/)]
9. Waljee AK, Higgins PD. Machine learning in medicine: a primer for physicians. *Am J Gastroenterol* 2010;105(6):1224-1226. [doi: [10.1038/ajg.2010.173](https://doi.org/10.1038/ajg.2010.173)] [Medline: [20523307](https://pubmed.ncbi.nlm.nih.gov/20523307/)]

## Abbreviations

**EDW:** enterprise data warehouse

**EHR:** electronic health record

**HELLP:** hemolysis, elevated liver enzymes, and low platelets

**HOPE-CAT:** Health Outcomes for all Pregnancy Experiences–Cardiovascular–Risk Assessment Technology

**ICD-10:** International Classification of Diseases, 10th Revision

**ML:** machine learning

**MMM:** maternal morbidity and mortality

*Edited by C Lovis; submitted 12.11.21; peer-reviewed by P Sarajlic; comments to author 03.12.21; revised version received 06.12.21; accepted 02.01.22; published 10.02.22.*

*Please cite as:*

*Shara N, Anderson KM, Falah N, Ahmad MF, Tavazoei D, Hughes JM, Talmadge B, Crovatt S, Dempers R*

*Early Identification of Maternal Cardiovascular Risk Through Sourcing and Preparing Electronic Health Record Data: Machine Learning Study*

*JMIR Med Inform 2022;10(2):e34932*

*URL: <https://medinform.jmir.org/2022/2/e34932>*

*doi: [10.2196/34932](https://doi.org/10.2196/34932)*

*PMID: [35142637](https://pubmed.ncbi.nlm.nih.gov/35142637/)*

©Nawar Shara, Kelley M Anderson, Noor Falah, Maryam F Ahmad, Darya Tavazoei, Justin M Hughes, Bethany Talmadge, Samantha Crovatt, Ramon Dempers. Originally published in JMIR Medical Informatics (<https://medinform.jmir.org>), 10.02.2022. This is an open-access article distributed under the terms of the Creative Commons Attribution License (<https://creativecommons.org/licenses/by/4.0/>), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Medical Informatics, is properly cited. The complete bibliographic information, a link to the original publication on <https://medinform.jmir.org/>, as well as this copyright and license information must be included.

Original Paper

# Development of Deep Learning Models for Predicting In-Hospital Mortality Using an Administrative Claims Database: Retrospective Cohort Study

Hiroki Matsui<sup>1</sup>, MPH; Hayato Yamana<sup>2</sup>, MD, PhD; Kiyohide Fushimi<sup>3</sup>, MD, PhD; Hideo Yasunaga<sup>1</sup>, MD, PhD

<sup>1</sup>Department of Clinical Epidemiology and Health Economics, School of Public Health, The University of Tokyo, Tokyo, Japan

<sup>2</sup>Department of Health Services Research, Graduate School of Medicine, The University of Tokyo, Tokyo, Japan

<sup>3</sup>Department of Health Policy and Informatics, Tokyo Medical and Dental University Graduate School, Tokyo, Japan

**Corresponding Author:**

Hiroki Matsui, MPH

Department of Clinical Epidemiology and Health Economics

School of Public Health

The University of Tokyo

7-3-1 Hongo

Bunkyo-ku

Tokyo, 1130033

Japan

Phone: 81 3 5841 1887

Fax: 81 3 5841 1888

Email: [ptmatsui-ky@umin.ac.jp](mailto:ptmatsui-ky@umin.ac.jp)

## Abstract

**Background:** Administrative claims databases have been used widely in studies because they have large sample sizes and are easily available. However, studies using administrative databases lack information on disease severity, so a risk adjustment method needs to be developed.

**Objective:** We aimed to develop and validate deep learning-based prediction models for in-hospital mortality of acute care patients.

**Methods:** The main model was developed using only administrative claims data (age, sex, diagnoses, and procedures on the day of admission). We also constructed disease-specific models for acute myocardial infarction, heart failure, stroke, and pneumonia using common severity indices for these diseases. Using the Japanese Diagnosis Procedure Combination data from July 2010 to March 2017, we identified 46,665,933 inpatients and divided them into derivation and validation cohorts in a ratio of 95:5. The main model was developed using a 9-layer deep neural network with 4 hidden dense layers that had 1000 nodes and were fully connected to adjacent layers. We evaluated model discrimination ability by an area under the receiver operating characteristic curve (AUC) and calibration ability by calibration plot.

**Results:** Among the eligible patients, 2,005,035 (4.3%) died. Discrimination and calibration of the models were satisfactory. The AUC of the main model in the validation cohort was 0.954 (95% CI 0.954-0.955). The main model had higher discrimination ability than the disease-specific models.

**Conclusions:** Our deep learning-based model using diagnoses and procedures produced valid predictions of in-hospital mortality.

(*JMIR Med Inform* 2022;10(2):e27936) doi:[10.2196/27936](https://doi.org/10.2196/27936)

## KEYWORDS

prognostic model; deep learning; real-world data; acute care; claims data; myocardial infarction; heart failure; stroke; pneumonia

## Introduction

Administrative claims databases have been used widely in clinical and epidemiological studies because they have large

sample sizes and are easily available. However, administrative data generally lack clinical information [1,2] and do not distinguish between comorbidities at admission and complications after admission [3]. Risk adjustment is not necessarily feasible in studies that use administrative databases

because of the lack of data on disease severity, and inadequate risk adjustment can result in confounding by indications.

Various models to predict in-hospital mortality have been developed using comorbidities recorded in administrative data. On the basis of these models, risk scores have been created and used to adjust for disease severity in clinical and epidemiological studies. However, the validity and usability of these models remain controversial [1,4-6]. For example, the Charlson comorbidity index was developed to predict in-hospital mortality, and is commonly used as a risk adjustment measure to capture levels of morbidity in studies that use administrative claims databases. However, this index only uses information on comorbidities that are recorded in the International Statistical Classification of Diseases and Related Health Problems, 10th revision (ICD-10) system [4].

Previous studies showed that additional clinical information improved the performance of mortality prediction models using administrative databases. In a previous study, we developed a procedure-based prediction model using the Japanese Diagnosis Procedure Combination (DPC) database, a nationwide administrative claims database [7]. However, these previous studies used logistic regression models that included only limited numbers of predictors.

Recent advances in machine learning (including deep learning) methods have made it possible to handle large amounts of information and complex models [8,9]. Machine learning methods allow researchers to input a large number of predictors, and variable selection is performed automatically. Conversely, conventional logistic regression requires variable selection based on the existing knowledge of experts.

Many previous studies have used machine learning to create disease-specific mortality prediction models (including models of heart failure [10], stroke [11], and myocardial infarction [12]), as well as all-patient mortality prediction models [12,13]. Most of these models used electronic health records and test results [9-13]. However, to collect such data from a wide range of medical institutions, it is necessary to standardize the electronic medical records. Furthermore, to use such data for clinical and epidemiological studies, experts in each disease area must manually extract information on predictor variables that are specific to the target disease. These factors make it difficult to standardize and use electronic medical records in a nationwide setting.

In this study, we developed and validated a deep learning-based model for predicting all-patient in-hospital mortality using only administrative claims data (including diagnoses and procedure data), which are uniformly formatted and routinely collected in a nationwide setting. To test the performance of the all-patient model, we also constructed disease-specific models for predicting in-hospital mortality of patients with acute myocardial infarction (AMI), heart failure (HF), stroke, or pneumonia, using common severity indices for each disease subgroup. Then, we compared the prediction abilities between the all-patient model and the disease-specific models for each disease subgroup.

## Methods

### Data Source

We conducted a retrospective cohort study. The data from July 2010 to March 2017 were collected from the DPC database. All the patients in the database were included to maximize the generalizability of the results. During the study period, 1569 hospitals contributed to the database. The patients in the database represented about 50% of all the acute-care inpatients in Japan [14].

The following data are included in the DPC database: age, sex, admission date, discharge date, diagnoses, and procedures (drugs, examinations, and surgical and nonsurgical treatments) for each patient. In the DPC database, comorbidities present at admission are clearly distinguished from complications arising after admission. All diagnoses were recorded using the International Statistical Classification of Diseases and Related Health Problems, 10th revision (ICD-10) codes. Procedure records were coded with Japanese conventional codes.

The DPC database also includes several severity indices, namely, the Killip classification for AMI [15,16], New York Heart Association classification for HF [17], Barthel index score for activities of daily living at admission [18], Japan Coma Scale of consciousness level at admission [19]; and age, dehydration, respiration, orientation, blood pressure (A-DROP), the Japan Respiratory Society community-acquired pneumonia severity index [20,21]. The Japan Coma Scale is used widely in Japan to measure impaired consciousness: a score of 0 indicates alert consciousness; single-digit scores (1, 2, 3) indicate being awake without stimuli; double-digit scores (10, 20, 30) indicate patients can be aroused by some stimuli; and triple-digit scores (100, 200, 300) indicate coma. A-DROP is a system for scoring severity of pneumonia that includes age (men  $\geq 70$  years, women  $\geq 75$  years), dehydration (serum urea nitrogen  $\geq 21$  mg/dL), respiratory failure (oxygen saturation by pulse oximetry  $\leq 90\%$  or PaO<sub>2</sub>  $\leq 60$  mm Hg), orientation disturbance (confusion), and low blood pressure (systolic blood pressure  $\leq 90$  mm Hg).

Our study was approved by the Ethics Committee of the University of Tokyo School of Medicine (approval number: 3501-(4)).

### Patient Selection

We extracted the data of inpatients who were discharged from hospitals between July 1, 2010, and March 31, 2017. The study population was divided randomly into a derivation cohort (95%) and a validation cohort (5%). For cases with 1-day hospitalization, the time at which we collected the information for prediction and the time at which the outcome occurred could be simultaneous. Because this could lead to an overestimation of the accuracy of the prediction model, we excluded patients who were discharged or died on the day of hospitalization from the validation cohort.

### Variables

The outcome variable was in-hospital death. For predictive variables, we used patients' demographic information (age, sex, and history of hospitalization in the 180 days before admission),

all the ICD-10-based diagnoses at admission, and all the procedures performed on the day of admission. Age was handled as a continuous variable; the other variables were handled as dichotomous variables (0 or 1). We also extracted the Killip classification [15,16], New York Heart Association classification [17], Barthel index score [18], Japan Coma Scale [19], and the A-DROP score [20,21] as common severity indices for specific diseases from the DPC database.

### Development of the Main Model

We developed a deep neural network model as the main model for predicting in-hospital death for all the patients, using 9 layers with 4 hidden dense layers [22,23]. For this, we used the patients' demographic information, all the ICD-10-based diagnoses at admission, and all the procedures performed on the day of admission. All the layers had 1000 nodes and were fully connected to adjacent layers. We used a softmax layer with 2 nodes as the output layer [24]. Because the numbers of deceased and alive patients were very different, we weighted the deceased cases with the reciprocal of the proportion of deceased cases (ie,  $1/0.045=22.3$ ) [23]. We used stochastic gradient descent to obtain neural network weights iteratively [25]. To avoid overfitting, 20% drop-out layers were sandwiched within each of the dense layers and an early stopping procedure involving learning steps using 3% data in the derivation cohort was employed [26]. Details of the weight optimization process are described in [Multimedia Appendix 1](#).

### Development of the Disease-Specific Models

We constructed disease-specific models for predicting in-hospital mortality in subgroups with AMI, HF, stroke, or pneumonia. The 4 models included patient backgrounds (age, sex, and history of hospitalization in the 180 days before admission) and diagnoses, and none of the models included procedures. For the AMI-specific model, we selected patients with AMI and included the Killip classification [15,16]. For the HF-specific model, we selected patients with HF and included the New York Heart Association classification [17]. For the stroke-specific model, we selected patients with stroke and included the Barthel index and the Japan Coma Scale at admission [18,19]. For the pneumonia-specific model, we selected patients with pneumonia and included the A-DROP score [20,21].

### Comparing Prediction Abilities Between the Main Model and the Disease-Specific Models

We applied the main model to the subgroups of patients with AMI, HF, stroke, and pneumonia and compared its prediction performance with the prediction performances of the disease-specific models for AMI, HF, stroke, and pneumonia.

We evaluated the performance of each model by calculating performance measures in the validation cohort. Performance measures included the area under the receiver operating characteristic curve (AUC), used to determine the discriminatory ability of the model. We calculated the 95% CI of the AUC using the DeLong method [27] and plotted a calibration curve to determine goodness of fit. We also calculated sensitivity, specificity, and positive and negative predictive values at the threshold determined by the Youden Index method [28]. We obtained CIs for all the indices with 2000 bootstraps.

We also examined whether the risk scores calculated by the disease-specific models improved the discrimination ability of the risk scores calculated by the main model. We incorporated the risk scores calculated by the main and disease-specific models into predictor variables of a logistic regression model and calculated combined risk scores that predicted in-hospital mortality for each disease population. The discrimination ability of the combined risk score was evaluated by its AUC and compared with the AUC of the main model. CIs for the AUC and hypothesis testing for the difference between the main model risk score and combined risk score were calculated using the DeLong method.

## Results

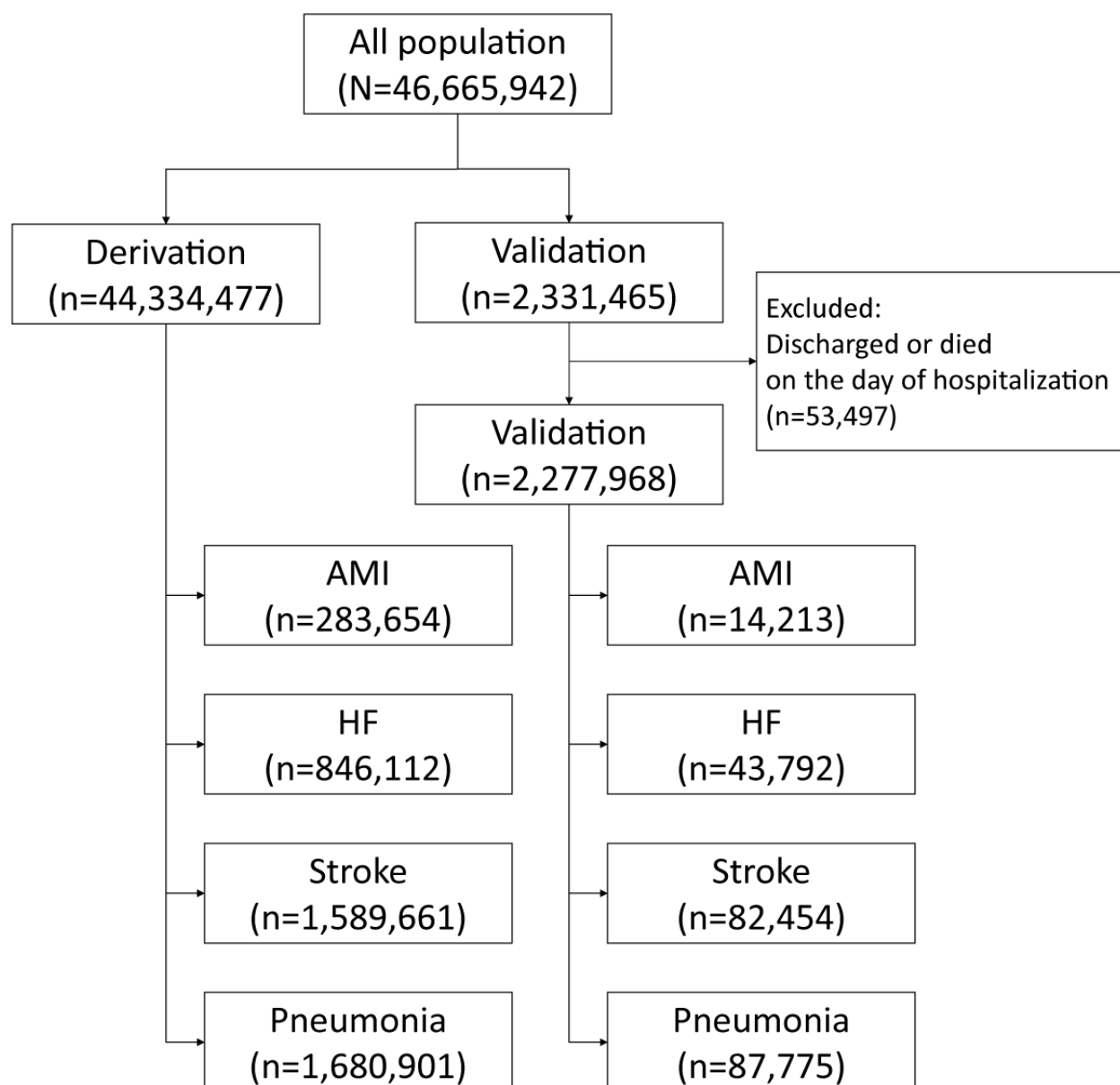
We obtained the data for 46,665,942 patients during the study period from the DPC database and divided them into derivation ( $n=44,334,477$ ) and validation ( $n=2,331,465$ ) cohorts. We excluded patients from the validation cohort according to the exclusion criteria, which left 2,277,968 patients in the validation cohort ([Figure 1](#)).

The characteristics of the derivation and validation cohorts are shown in [Table 1](#). The average lengths of stay were 14.2 days and 14.5 days and in-hospital mortality was 4.3% and 3.7% in the derivation and validation cohorts, respectively. Patients in the validation cohort were slightly older and had more comorbidities than those in the derivation cohort.

The structure of the main model is shown in [Table 2](#). There were 49,297 predictor variables, including 3 variables on patient demographics and history (age, sex, history of hospitalization in the 180 days before admission), 19,930 diagnoses at admission, and 29,364 procedures (drugs, examinations, surgical and nonsurgical treatments). Overall, 52,302,002 weights ( $=49,297 \times 1000 + 1001 \times 1000 + 1001 \times 1000 + 1001 \times 1000 + 1001 \times 2$ ) of links between the layers were optimized in the derivation. The script for the deep learning model including model weights is available on our website [29].



**Figure 1.** Numbers of patients in the derivation and validation cohorts and disease-specific subgroups. AMI: acute myocardial infarction, HF: heart failure.



**Table 1.** Characteristics of the patients in the derivation and validation cohorts.

Characteristic	Derivation cohort (n=44,334,477)	Validation cohort (n=2,277,968)	P value
Death, n (%)	1,905,286 (4.3)	83,292 (3.7)	<.001
Length of hospital stay (days), mean (SD)	14.2 (24.1)	14.5 (24.2)	<.001
Age (years), mean (SD)	60.1 (24.4)	60.4 (24.2)	<.001
Sex (male), n (%)	23,480,628 (53.0)	1,207,886 (53.0)	.07
History of hospitalization within 180 days, n (%)	12,282,386 (27.7)	632,362 (27.8)	.07
<b>Charlson comorbidity index, n (%)</b>			<.001
0-1	28,734,890 (64.8)	1,465,779 (64.3)	
2-3	11,432,403 (25.8)	594,500 (26.1)	
≥4	4,165,579 (9.4)	217,605 (9.6)	

**Table 2.** Structure of the main model.

Layer	Input (nodes)	Output (nodes)	Weights, n
1: Input	49,297	1000	49,297,000
2: Drop-out	N/A <sup>a</sup>	N/A	N/A
3: Hidden 1	1001	1000	1,001,000
4: Drop-out	N/A	N/A	N/A
5: Hidden 2	1001	1000	1,001,000
6: Drop-out	N/A	N/A	N/A
7: Hidden 3	1001	1000	1,001,000
8: Drop-out	N/A	N/A	N/A
9: Output	1001	2	2002
Sum of weights	N/A	N/A	52,302,002

<sup>a</sup>N/A: not applicable.

An overview of the main and disease-specific models used in this study is given in [Table 3](#). The total number of weights was calculated as follows: total number of weights = the number of input nodes  $\times$  1000 + 1001  $\times$  1000 + 1001  $\times$  1000 + 1001  $\times$  1000 + 1001  $\times$  2.

The AUC of the main model in the validation cohort was 0.954 (95% CI 0.9537-0.9547). The sensitivity, specificity, and positive and negative predictive values at the cutoff point (0.0435) determined by the Youden index method of the main model were 0.920 (95% CI 0.915-0.924), 0.855 (95% CI 0.852-0.860), 0.195 (95% CI 0.192-0.199), and 0.996 (95% CI 0.996-0.997), respectively ([Table 4](#)).

The calibration curves of the observed and estimated mortality in the validation cohort are shown in [Figure 2](#). Observed and estimated mortality were strongly correlated, but the estimated mortality was slightly lower than the observed mortality.

The AUCs and other prediction metrics of the main and disease-specific models are shown in [Table 4](#). The AUCs of the main model for the AMI, HF, stroke, and pneumonia subgroups were 0.944, 0.832, 0.921, and 0.918, respectively. The AUCs of the disease-specific models for the AMI, HF, stroke, and pneumonia subgroups were 0.876, 0.745, 0.894, and 0.863, respectively. The main model showed significantly higher discriminant ability than the disease-specific models for all 4 subgroups.

**Table 3.** Summary of the main and disease-specific models.

Model	Input (nodes)	Weights, N
Main model	49,297	52,302,002
Acute myocardial infarction model	9	3,014,002
Stroke model	54	3,059,002
Heart failure model	9	3,014,002
Pneumonia model	9	3,014,002

**Table 4.** Performances of the main and disease-specific models.

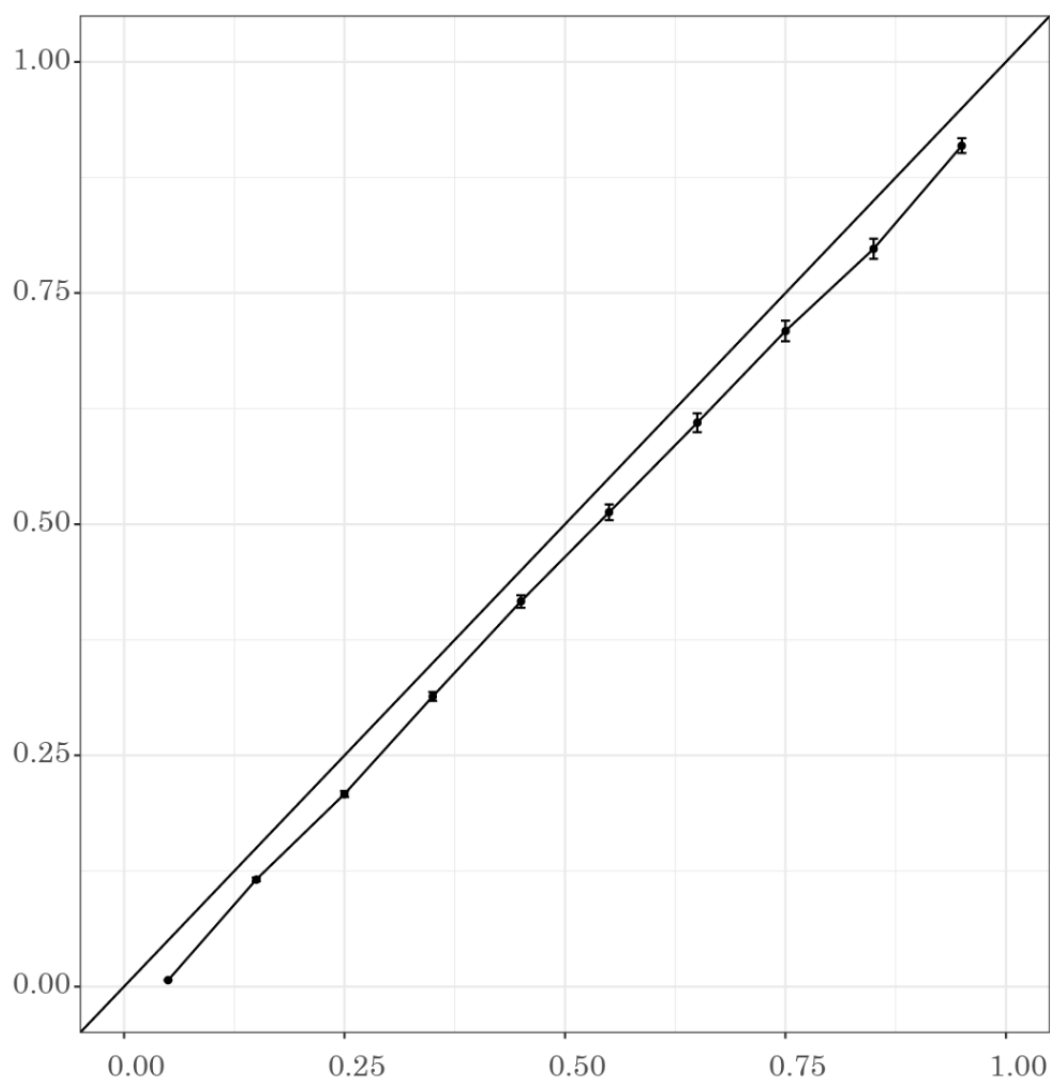
	AUC <sup>a</sup> (95% CI)	Threshold	Sensitivity (95% CI)	Specificity (95% CI)	PPV <sup>b</sup> (95% CI)	NPV <sup>c</sup> (95% CI)
<b>Validation cohort (n=2,331,465)</b>						
Main model	0.954 (0.954-0.955)	0.0435	0.920 (0.915-0.924)	0.855 (0.852-0.860)	0.195 (0.192-0.199)	0.996 (0.996-0.997)
<b>Acute myocardial infarction (n=14,213)</b>						
Main model	0.944 (0.938-0.950)	0.087	0.888 (0.864-0.947)	0.862 (0.796-0.881)	0.334 (0.264-0.363)	0.990 (0.988-0.995)
Disease-specific model	0.876 (0.866-0.887)	0.087	0.837 (0.797-0.877)	0.783 (0.745-0.817)	0.233 (0.210-0.257)	0.984 (0.981-0.988)
<b>Heart failure (n=43,792)</b>						
Main model	0.831 (0.825-0.837)	0.118	0.782 (0.729-0.813)	0.719 (0.678-0.771)	0.220 (0.205-0.245)	0.970 (0.965-0.973)
Disease-specific model	0.745 (0.738-0.753)	0.097	0.727 (0.678-0.754)	0.642 (0.613-0.688)	0.172 (0.166-0.184)	0.958 (0.954-0.961)
<b>Stroke (n=82,454)</b>						
Main model	0.921 (0.918-0.925)	0.091	0.863 (0.847-0.901)	0.824 (0.781-0.837)	0.267 (0.234-0.279)	0.988 (0.987-0.991)
Disease-specific model	0.894 (0.890-0.898)	0.080	0.824 (0.805-0.836)	0.800 (0.793-0.818)	0.235 (0.229-0.249)	0.984 (0.983-0.985)
<b>Pneumonia (n=87,775)</b>						
Main model	0.918 (0.915-0.920)	0.075	0.913 (0.896-0.925)	0.769 (0.762-0.786)	0.209 (0.204-0.219)	0.993 (0.991-0.994)
Disease-specific model	0.863 (0.859-0.867)	0.064	0.851 (0.809-0.913)	0.705 (0.638-0.744)	0.160 (0.143-0.173)	0.986 (0.983-0.991)

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

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

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

**Figure 2.** Calibration curves for the observed and estimated mortality in the validation cohort with the main model. X-axis indicates predicted mortality and Y-axis indicates actual mortality.



The discriminatory ability of the combined risk scores and the risk scores calculated by the main model are shown in [Table 5](#). All combined risk scores except the one for AMI had significantly higher AUCs than the main model risk scores. However, the differences between the main model risk scores and the combined risk scores were small.

The calibration curves for the main and disease-specific models for the subgroups are shown in [Figure 3](#). The correlations

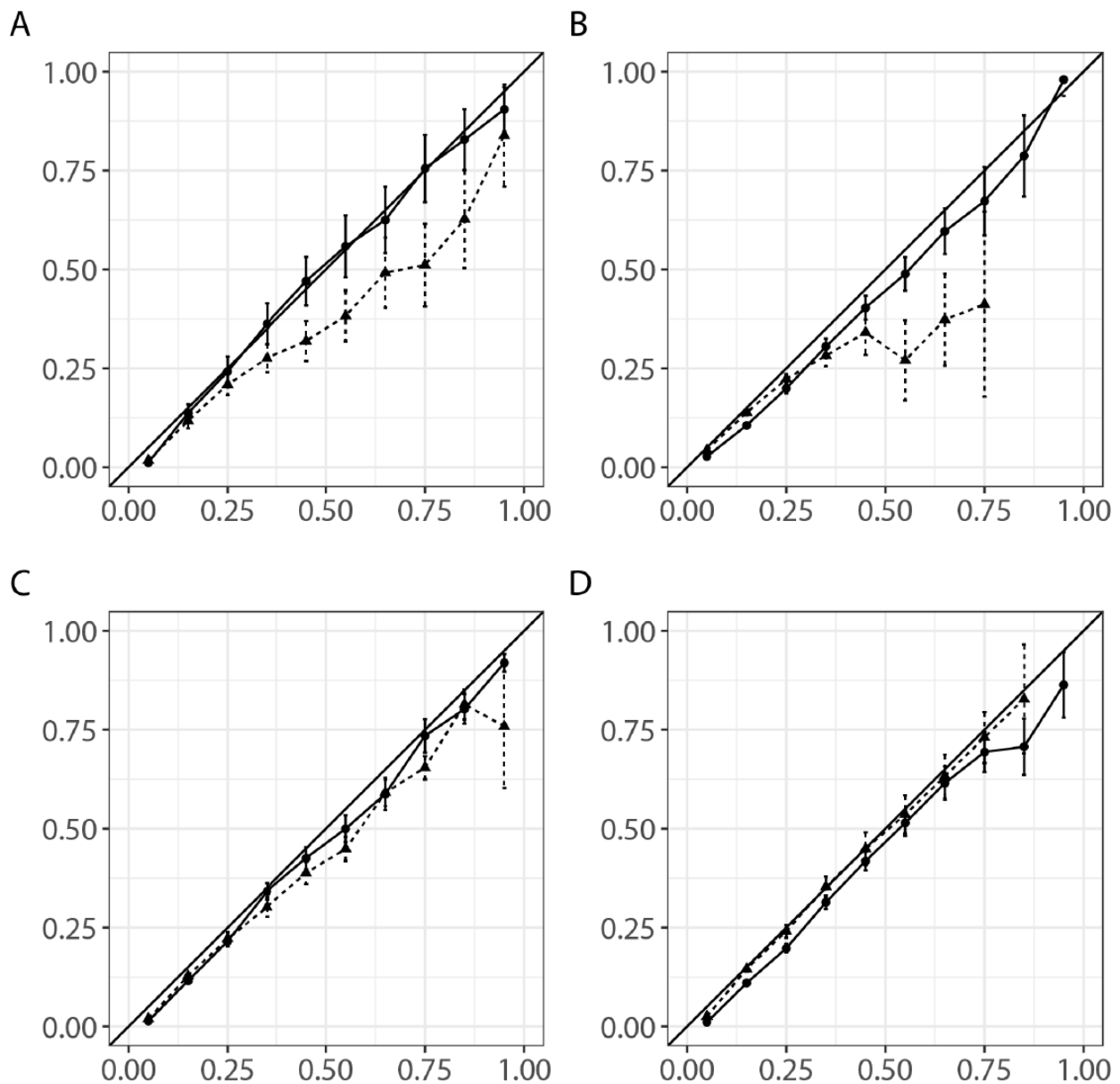
between the observed and estimated mortality were better with the main model than with the disease-specific models for the AMI, HF, and stroke subgroups ([Figure 3A-C](#)). For the pneumonia subgroup, the correlations were similar between the main and disease-specific models when the predicted mortality was  $\leq 0.8$ . However, the disease-specific model failed to estimate mortality well when the predicted mortality was  $\geq 0.8$  ([Figure 3D](#)).

**Table 5.** Comparison of the discriminatory ability of the combined risk scores and the risk scores calculated by the main model.

	Main model AUC <sup>a</sup> (95% CI)	Combined risk score AUC (95% CI)	P value
Acute myocardial infarction	0.944 (0.938-0.950)	0.945 (0.939-0.951)	.23
Heart failure	0.831 (0.825-0.837)	0.838 (0.832-0.844)	<.001
Stroke	0.921 (0.918-0.925)	0.927 (0.924-0.930)	<.001
Pneumonia	0.918 (0.915-0.920)	0.921 (0.918-0.924)	<.001

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

**Figure 3.** Calibration curves for the observed and estimated mortality in the validation cohort with the disease-specific models. Models for (A) acute myocardial infarction, (B) heart failure, (C) stroke, and (D) pneumonia. X-axis: predicted mortality. Y-axis: actual mortality. Solid line: main model. Dotted line: disease-specific models.



## Discussion

### Principal Findings

We constructed deep learning-based prediction models for in-hospital mortality using a large Japanese inpatient database. Patient backgrounds, diagnoses, and treatments on the first day of admission were entered into the models. The overall discriminant abilities of the models were high in subgroups of patients with AMI, HF, stroke, and pneumonia. The main model had better discriminant abilities than disease-specific models using common severity indices. We integrated the risk scores for the main and disease-specific models and calculated combined risk scores. However, the improvement in the predictive performance of the combined risk scores over that of the main model risk scores was only slight.

Risk scores derived from administrative claims databases have been developed previously. For example, the Charlson and Elixhauser models, which use comorbidity information to predict long-term survival, have been used for risk adjustment in clinical and epidemiological studies [30,31]. In this study, a new prediction model for in-hospital mortality developed using administrative claims data showed high discriminatory power (AUC=0.945). We believe that our model can also be used for risk adjustment in clinical and epidemiological studies using administrative claims data that includes diagnoses and procedures.

In a previous study, we constructed a prediction model for in-hospital mortality that incorporated comorbidities and several selected procedures (blood tests, radiography, echocardiogram) on the day of admission [7]. However, that model lacked

generalizability; for example, it was not applicable to critically ill patients. The newly constructed model can be used for risk prediction and adjustment for patients with a wide range of disease severity.

In a previous study, the predictive abilities of models with administrative claims data alone were compared with those of models with electronic medical records combined with administrative claims data [32]. The predictive abilities of the models with electronic medical records were higher because the electronic medical records included detailed information related to each patient, such as blood test results, vital signs, and admission data collected during the first 2 days of the index admission.

In this study, a deep learning model that used only massive administrative data had higher predictive ability than models that used disease-specific severity information. On the basis of our results, we consider that large-scale administrative data can be used to predict in-hospital mortality more accurately than the generally used severity indices. Kharrazi et al [33] reported that obtaining information from both administrative data and electronic health records increased the prediction accuracy of their model compared with using each data source alone. Zeltzer et al [32] found that feeding the electronic health record information collected during hospitalization, in addition to the administrative data and pre-hospitalization electronic health record information, into their model resulted in more accurate mortality risk assessment. Rajkomar et al [9] predicted in-hospital mortality with the same level of accuracy as we achieved in this study by using information from structured electronic health records. We also found that a combined risk score, obtained by integrating the main model with a disease-specific model, showed higher prediction accuracy than the risk score obtained from the main model. However, in this study, the difference between the main model and the combined risk score was small, and there was no significant difference between the two risk scores for AMI. This indicates that the

main model was able to construct a risk score comparable to the combined risk score even without disease-specific severity information. Therefore, we propose that patient outcome studies can be conducted using administrative data alone, such as the initial hospitalization process and diagnosis, without the need for data on disease severity.

It is not easy to collect electronic health record information in a standardized way and use it for research. We believe that the results of this study can be used in cases where it is not possible to obtain detailed clinical information, such as disease severity and vital signs, that would be included in an electronic health record.

### Limitations

This study has several limitations. First, we did not conduct an external validation. Second, we did not use a variety of machine learning methods (eg, random forest, lasso regression, XGBoost, and their ensembles), so we could not compare the prediction performance of other machine learning methods. Third, because the database used in this study is for acute hospitalization, we could not obtain data on long-term outcomes. Fourth, model accuracy is not always guaranteed for all diseases, so the applicability of the model to other populations needs to be considered.

### Conclusion

In conclusion, we constructed a deep neural network model to predict in-hospital mortality using all the data on diagnoses and procedures performed on the day of admission in a Japanese administrative claims database. Our model using only administrative claims data showed higher prediction ability than our models using the more generally used severity indices. We propose that prognostic models using data on diagnoses and procedures obtained only from administrative claims databases can predict in-hospital mortality and can be used for risk adjustment in clinical and epidemiological studies.

---

### Conflicts of Interest

None declared.

---

### Multimedia Appendix 1

Model weight optimization process details.

[\[PDF File \(Adobe PDF File\), 50 KB - medinform\\_v10i2e27936\\_app1.pdf\]](#)

---

### References

1. Sung S, Hsieh C, Kao Yang Y, Lin H, Chen C, Chen Y, et al. Developing a stroke severity index based on administrative data was feasible using data mining techniques. *J Clin Epidemiol* 2015 Nov;68(11):1292-1300. [doi: [10.1016/j.jclinepi.2015.01.009](https://doi.org/10.1016/j.jclinepi.2015.01.009)] [Medline: [25700940](https://pubmed.ncbi.nlm.nih.gov/25700940/)]
2. Virnig BA, McBean M. Administrative data for public health surveillance and planning. *Annu Rev Public Health* 2001;22:213-230. [doi: [10.1146/annurev.publhealth.22.1.213](https://doi.org/10.1146/annurev.publhealth.22.1.213)] [Medline: [11274519](https://pubmed.ncbi.nlm.nih.gov/11274519/)]
3. Yamana H, Matsui H, Sasabuchi Y, Fushimi K, Yasunaga H. Categorized diagnoses and procedure records in an administrative database improved mortality prediction. *J Clin Epidemiol* 2015 Sep;68(9):1028-1035. [doi: [10.1016/j.jclinepi.2014.12.004](https://doi.org/10.1016/j.jclinepi.2014.12.004)] [Medline: [25596112](https://pubmed.ncbi.nlm.nih.gov/25596112/)]
4. Sundararajan V, Quan H, Halfon P, Fushimi K, Luthi J, Burnand B, International Methodology Consortium for Coded Health Information (IMECCHI). Cross-national comparative performance of three versions of the ICD-10 Charlson index. *Med Care* 2007 Dec;45(12):1210-1215. [doi: [10.1097/MLR.0b013e3181484347](https://doi.org/10.1097/MLR.0b013e3181484347)] [Medline: [18007172](https://pubmed.ncbi.nlm.nih.gov/18007172/)]

5. Elixhauser A, Steiner C, Harris DR, Coffey RM. Comorbidity measures for use with administrative data. *Med Care* 1998 Jan;36(1):8-27. [doi: [10.1097/00005650-199801000-00004](https://doi.org/10.1097/00005650-199801000-00004)] [Medline: [9431328](https://pubmed.ncbi.nlm.nih.gov/9431328/)]
6. Pine M, Jordan HS, Elixhauser A, Fry DE, Hoaglin DC, Jones B, et al. Enhancement of claims data to improve risk adjustment of hospital mortality. *JAMA* 2007 Jan 03;297(1):71-76. [doi: [10.1001/jama.297.1.71](https://doi.org/10.1001/jama.297.1.71)] [Medline: [17200477](https://pubmed.ncbi.nlm.nih.gov/17200477/)]
7. Yamana H, Matsui H, Fushimi K, Yasunaga H. Procedure-based severity index for inpatients: development and validation using administrative database. *BMC Health Serv Res* 2015 Jul 08;15:261 [FREE Full text] [doi: [10.1186/s12913-015-0889-x](https://doi.org/10.1186/s12913-015-0889-x)] [Medline: [26152112](https://pubmed.ncbi.nlm.nih.gov/26152112/)]
8. Purushotham S, Meng C, Che Z, Liu Y. Benchmarking deep learning models on large healthcare datasets. *J Biomed Inform* 2018 Jul;83:112-134 [FREE Full text] [doi: [10.1016/j.jbi.2018.04.007](https://doi.org/10.1016/j.jbi.2018.04.007)] [Medline: [29879470](https://pubmed.ncbi.nlm.nih.gov/29879470/)]
9. 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(1):18 [FREE Full text] [doi: [10.1038/s41746-018-0029-1](https://doi.org/10.1038/s41746-018-0029-1)] [Medline: [31304302](https://pubmed.ncbi.nlm.nih.gov/31304302/)]
10. Rahimi K, Bennett D, Conrad N, Williams TM, Basu J, Dwight J, et al. Risk prediction in patients with heart failure: a systematic review and analysis. *JACC Heart Fail* 2014 Oct;2(5):440-446 [FREE Full text] [doi: [10.1016/j.jchf.2014.04.008](https://doi.org/10.1016/j.jchf.2014.04.008)] [Medline: [25194291](https://pubmed.ncbi.nlm.nih.gov/25194291/)]
11. 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](https://doi.org/10.1371/journal.pone.0234722)] [Medline: [32530947](https://pubmed.ncbi.nlm.nih.gov/32530947/)]
12. Cho S, Austin P, Ross H, Abdel-Qadir H, Chicco D, Tomlinson G, et al. Machine Learning Compared With Conventional Statistical Models for Predicting Myocardial Infarction Readmission and Mortality: A Systematic Review. *Can J Cardiol* 2021 Aug;37(8):1207-1214. [doi: [10.1016/j.cjca.2021.02.020](https://doi.org/10.1016/j.cjca.2021.02.020)] [Medline: [33677098](https://pubmed.ncbi.nlm.nih.gov/33677098/)]
13. Tabak YP, Sun X, Nunez CM, Johannes RS. Using electronic health record data to develop inpatient mortality predictive model: Acute Laboratory Risk of Mortality Score (ALaRMS). *J Am Med Inform Assoc* 2014;21(3):455-463 [FREE Full text] [doi: [10.1136/amiainl-2013-001790](https://doi.org/10.1136/amiainl-2013-001790)] [Medline: [24097807](https://pubmed.ncbi.nlm.nih.gov/24097807/)]
14. Yasunaga H. Real World Data in Japan: Chapter II The Diagnosis Procedure Combination Database. *ACE* 2019;1(3):76-79. [doi: [10.37737/ace.1.3\\_76](https://doi.org/10.37737/ace.1.3_76)]
15. Killip T, Kimball JT. Treatment of myocardial infarction in a coronary care unit. A two year experience with 250 patients. *Am J Cardiol* 1967 Oct;20(4):457-464. [doi: [10.1016/0002-9149\(67\)90023-9](https://doi.org/10.1016/0002-9149(67)90023-9)] [Medline: [6059183](https://pubmed.ncbi.nlm.nih.gov/6059183/)]
16. Shiraiishi J, Kohno Y, Nakamura T, Yanagiuchi T, Hashimoto S, Ito D, AMI-Kyoto Multi-Center Risk Study Group. Predictors of in-hospital outcomes after primary percutaneous coronary intervention for acute myocardial infarction in patients with a high Killip class. *Intern Med* 2014;53(9):933-939 [FREE Full text] [doi: [10.2169/internalmedicine.53.1144](https://doi.org/10.2169/internalmedicine.53.1144)] [Medline: [24785883](https://pubmed.ncbi.nlm.nih.gov/24785883/)]
17. White PD, Myers MM. The classification of cardiac diagnosis. *JAMA* 1921 Oct 29;77(18):1414. [doi: [10.1001/jama.1921.02630440034013](https://doi.org/10.1001/jama.1921.02630440034013)]
18. Duffy L, Gajree S, Langhorne P, Stott DJ, Quinn TJ. Reliability (inter-rater agreement) of the Barthel Index for assessment of stroke survivors: systematic review and meta-analysis. *Stroke* 2013 Mar;44(2):462-468 [FREE Full text] [doi: [10.1161/STROKEAHA.112.678615](https://doi.org/10.1161/STROKEAHA.112.678615)] [Medline: [23299497](https://pubmed.ncbi.nlm.nih.gov/23299497/)]
19. Shigematsu K, Nakano H, Watanabe Y. The eye response test alone is sufficient to predict stroke outcome—reintroduction of Japan Coma Scale: a cohort study. *BMJ Open* 2013;3(4):e002736 [FREE Full text] [doi: [10.1136/bmjopen-2013-002736](https://doi.org/10.1136/bmjopen-2013-002736)] [Medline: [23633419](https://pubmed.ncbi.nlm.nih.gov/23633419/)]
20. Miyashita N, Matsushima T, Oka M, Japanese Respiratory Society. The JRS guidelines for the management of community-acquired pneumonia in adults: an update and new recommendations. *Intern Med* 2006;45(7):419-428 [FREE Full text] [doi: [10.2169/internalmedicine.45.1691](https://doi.org/10.2169/internalmedicine.45.1691)] [Medline: [16679695](https://pubmed.ncbi.nlm.nih.gov/16679695/)]
21. Ahn JH, Choi EY. Expanded A-DROP Score: A New Scoring System for the Prediction of Mortality in Hospitalized Patients with Community-acquired Pneumonia. *Sci Rep* 2018 Oct 01;8(1):14588 [FREE Full text] [doi: [10.1038/s41598-018-32750-2](https://doi.org/10.1038/s41598-018-32750-2)] [Medline: [30275523](https://pubmed.ncbi.nlm.nih.gov/30275523/)]
22. Chollet F. User experience design for APIs. *The Keras Blog*. 2015. URL: <https://blog.keras.io/author/francois-chollet.html> [accessed 2022-02-04]
23. Abadi M, Agarwal A, Barham P, Brevdo E, Chen Z, Citro C, et al. TensorFlow: Large-Scale Machine Learning on Heterogeneous Distributed Systems. *arXiv. Preprint posted online March 14, 2016* 2016 [FREE Full text]
24. Geron A. Softmax Regression. In: *Hands-on machine learning with scikit-learn and tensorflow: Concepts Tools, and Techniques to Build Intelligent Systems*. 1st ed. Sebastopol, CA: O'Reilly Media; 2017:147-149.
25. Geron A. Stochastic Gradient Descent. In: *Hands-on machine learning with scikit-learn and tensorflow: Concepts Tools, and Techniques to Build Intelligent Systems*. 1st ed. Sebastopol, CA: O'Reilly Media; 2017:122-126.
26. Geron A. Avoiding Overfitting Through Regularization. In: *Hands-on machine learning with scikit-learn and tensorflow: Concepts Tools, and Techniques to Build Intelligent Systems*. 1st ed. Sebastopol, CA: O'Reilly Media; 2017:363-371.
27. DeLong ER, DeLong DM, Clarke-Pearson DL. Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. *Biometrics* 1988 Oct;44(3):837-845. [Medline: [3203132](https://pubmed.ncbi.nlm.nih.gov/3203132/)]
28. Youden WJ. Index for rating diagnostic tests. *Cancer* 1950 Jan;3(1):32-35. [doi: [10.1002/1097-0142\(1950\)3:1<32::aid-cnrc2820030106>3.0.co;2-3](https://doi.org/10.1002/1097-0142(1950)3:1<32::aid-cnrc2820030106>3.0.co;2-3)] [Medline: [15405679](https://pubmed.ncbi.nlm.nih.gov/15405679/)]

29. Matsui H. Development of deep learning models for predicting in-hospital mortality using an administrative claims database. Researchmap. URL: <https://researchmap.jp/ptmatsui> [accessed 2022-02-01]
30. Matsui H, Jo T, Fushimi K, Yasunaga H. Outcomes after early and delayed rehabilitation for exacerbation of chronic obstructive pulmonary disease: a nationwide retrospective cohort study in Japan. *Respir Res* 2017 Apr 21;18(1):68 [FREE Full text] [doi: [10.1186/s12931-017-0552-7](https://doi.org/10.1186/s12931-017-0552-7)] [Medline: [28431501](https://pubmed.ncbi.nlm.nih.gov/28431501/)]
31. Matsui H, Koike S, Fushimi K, Wada T, Yasunaga H. Effect of neurologic specialist staffing on 30-day in-hospital mortality after cerebral infarction. *ACE* 2019;1(3):86-94. [doi: [10.37737/ace.1.3.86](https://doi.org/10.37737/ace.1.3.86)]
32. Zeltzer D, Balicer RD, Shir T, Flaks-Manov N, Einav L, Shadmi E. Prediction Accuracy With Electronic Medical Records Versus Administrative Claims. *Med Care* 2019 Jul;57(7):551-559. [doi: [10.1097/MLR.0000000000001135](https://doi.org/10.1097/MLR.0000000000001135)] [Medline: [31135691](https://pubmed.ncbi.nlm.nih.gov/31135691/)]
33. 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](https://doi.org/10.1097/MLR.0000000000000754)] [Medline: [28598890](https://pubmed.ncbi.nlm.nih.gov/28598890/)]

## Abbreviations

**A-DROP:** age, dehydration, respiratory failure, orientation disturbance, and low blood pressure

**AMI:** acute myocardial infarction

**AUC:** area under the receiver operating characteristic curve

**DPC:** Japanese Diagnosis Procedure Combination

**HF:** heart failure

**ICD-10:** International Statistical Classification of Diseases and Related Health Problems, 10th revision

**NPV:** negative predictive value

**PPV:** positive predictive value

*Edited by C Lovis; submitted 13.02.21; peer-reviewed by E Shadmi, N Flaks-Manov; comments to author 02.04.21; revised version received 05.06.21; accepted 02.01.22; published 11.02.22.*

*Please cite as:*

*Matsui H, Yamana H, Fushimi K, Yasunaga H*

*Development of Deep Learning Models for Predicting In-Hospital Mortality Using an Administrative Claims Database: Retrospective Cohort Study*

*JMIR Med Inform* 2022;10(2):e27936

URL: <https://medinform.jmir.org/2022/2/e27936>

doi: [10.2196/27936](https://doi.org/10.2196/27936)

PMID: [34997958](https://pubmed.ncbi.nlm.nih.gov/34997958/)

©Hiroki Matsui, Hayato Yamana, Kiyohide Fushimi, Hideo Yasunaga. Originally published in JMIR Medical Informatics (<https://medinform.jmir.org>), 11.02.2022. This is an open-access article distributed under the terms of the Creative Commons Attribution License (<https://creativecommons.org/licenses/by/4.0/>), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Medical Informatics, is properly cited. The complete bibliographic information, a link to the original publication on <https://medinform.jmir.org/>, as well as this copyright and license information must be included.



Original Paper

# Panic Attack Prediction Using Wearable Devices and Machine Learning: Development and Cohort Study

Chan-Hen Tsai<sup>1,2</sup>, MD; Pei-Chen Chen<sup>1</sup>, MSc; Ding-Shan Liu<sup>3</sup>, BSc; Ying-Ying Kuo<sup>2</sup>, MSc; Tsung-Ting Hsieh<sup>1</sup>, BSc; Dai-Lun Chiang<sup>4</sup>, PhD; Feipei Lai<sup>1,3</sup>, PhD; Chia-Tung Wu<sup>3</sup>, MSc

<sup>1</sup>Graduate Institute of Biomedical Electronics and Bioinformatics, National Taiwan University, Taipei City, Taiwan

<sup>2</sup>Department of Psychiatry, En Chu Kong Hospital, New Taipei City, Taiwan

<sup>3</sup>Department of Computer Science and Information Engineering, National Taiwan University, Taipei City, Taiwan

<sup>4</sup>Financial Technology Applications Program, Ming Chuan University, Taoyuan City, Taiwan

**Corresponding Author:**

Chia-Tung Wu, MSc

Department of Computer Science and Information Engineering

National Taiwan University

No. 1, Sec. 4, Roosevelt Rd.

Taipei City, 106319

Taiwan

Phone: 886 978006469

Email: [tony006469@gmail.com](mailto:tony006469@gmail.com)

## Abstract

**Background:** A panic attack (PA) is an intense form of anxiety accompanied by multiple somatic presentations, leading to frequent emergency department visits and impairing the quality of life. A prediction model for PAs could help clinicians and patients monitor, control, and carry out early intervention for recurrent PAs, enabling more personalized treatment for panic disorder (PD).

**Objective:** This study aims to provide a 7-day PA prediction model and determine the relationship between a future PA and various features, including physiological factors, anxiety and depressive factors, and the air quality index (AQI).

**Methods:** We enrolled 59 participants with PD (*Diagnostic and Statistical Manual of Mental Disorders, 5th edition*, and the Mini International Neuropsychiatric Interview). Participants used smartwatches (Garmin Vivosmart 4) and mobile apps to collect their sleep, heart rate (HR), activity level, anxiety, and depression scores (Beck Depression Inventory [BDI], Beck Anxiety Inventory [BAI], State-Trait Anxiety Inventory state anxiety [STAI-S], State-Trait Anxiety Inventory trait anxiety [STAI-T], and Panic Disorder Severity Scale Self-Report) in their real life for a duration of 1 year. We also included AQIs from open data. To analyze these data, our team used 6 machine learning methods: random forests, decision trees, linear discriminant analysis, adaptive boosting, extreme gradient boosting, and regularized greedy forests.

**Results:** For 7-day PA predictions, the random forest produced the best prediction rate. Overall, the accuracy of the test set was 67.4%-81.3% for different machine learning algorithms. The most critical variables in the model were questionnaire and physiological features, such as the BAI, BDI, STAI, MINI, average HR, resting HR, and deep sleep duration.

**Conclusions:** It is possible to predict PAs using a combination of data from questionnaires and physiological and environmental data.

(*JMIR Med Inform* 2022;10(2):e33063) doi:[10.2196/33063](https://doi.org/10.2196/33063)

**KEYWORDS**

panic disorder; panic attack; prediction; wearable device; machine learning; lifestyle

## Introduction

### Background

Panic disorder (PD) is a common mental disorder with a lifetime prevalence of about 1.6%-3.5% worldwide [1,2]. Its main characteristic is the fear of recurrent panic attacks (PAs) and loss of control, which leads to functional impairment. Patients suffering from PD often make frequent visits to the emergency department before formal diagnosis and psychoeducation. Functional impairment of PD can be avoidant behavior in terms of crowds, open spaces, traffic vehicles, or stressful situations. Severe PD cases [3] may become homebound. Accurate PA prediction may help clinicians to provide appropriate, timely treatment and to optimize personalized medicine.

A PA is typically an abrupt surge of intense fear reaching a peak within minutes, including 4 or more of the following symptoms: palpitations; sweating; trembling or shaking; sensations of shortness of breath or smothering; a feeling of choking; chest tightness; nausea or abdominal distress; dizziness or faintness; derealization (feelings of unreality) or depersonalization (being detached from oneself); fear of losing control, or going crazy; fear of dying; numbness or tingling sensation; chills; and heat sensation disturbance. A PA with fewer than 4 symptoms is called a limited panic attack (limited PA). Due to its high prevalence, the *Diagnostic and Statistical Manual of Mental Disorders, 5th edition* (DSM-5) [4], uses PA as a descriptive specifier across all mental disorders [5].

### Theory and Hypothesis

PAs are known to be triggered by psychological stress or specific occasions that induce a fear of being unable to escape (agoraphobia). However, so far, few studies have predicted recurrent PAs using real-life data. We hypothesize that recurrent PAs are related to multiple factors, including physiological, emotional, and personality factors. Cho et al [6] and Trushna et al [7] further observed a positive association between PAs and air pollution. To confirm these associations, we evaluated PAs from various domains to establish a more explainable model.

### Previous Work

Researchers have used a variety of data sources to predict PD severity and prognosis, including demographic features, clinical scales, diagnostic information, medical history, functional magnetic resonance imaging (fMRI), electrocardiogram (ECG), electroencephalogram (EEG), and genetic data, such as DNA methylation signatures [8]. In recent years, a few studies have begun to use watch-type computers, wearable devices, or physical challenge by CO<sub>2</sub> [9] to predict PAs. However, there is no clear evidence showing which features are superior for prediction. It is also difficult to compare these studies due to the heterogeneity of study design, methods, and sample selections. Next, we give a brief review.

*Clinical questionnaires* with internal consistency and reliability are the tools widely used to predict PA and PD severity. These tools assess the participant's emotional and personality traits, for example, the Anxiety Sensitivity Index (ASI) [10,11], the State-Trait Personality Inventory (STPI) [12], Hamilton

Depression Rating Scale (HAM-D), Beck Anxiety Inventory (BAI), and State-Trait Anxiety Inventory (STAI). Liu et al [13] used 11 predictors for PD recurrence from past demographic, clinical, and psychosocial factors, yielding a discriminative power C statistic of 72.8%. Most clinical questionnaires can be delivered at clinics or via internet-based approaches.

*fMRI* [14-17] compares areas of brain activation before and after a particular treatment, clarifies the structural change in PD, and predicts whether PD is comorbid with depression. However, predictions from different fMRI studies are inconsistent [18]. In addition, fMRI is expensive and complicates real-time PA prediction. It is, instead, an excellent tool by which to explore the psychopathology of PD.

An *EEG* detects specific patterns, such as slow waves in the  $\theta$ -band, in PD patients, as shown in a study [19]. A review of EEG [20] summarized that PD tends to show decreased  $\alpha$ -band power and increased  $\beta$ -band power, but the review did not yield an algorithm to predict PAs using EEG patterns.

*Wearable devices* are the most promising tool by which to detect PAs throughout the patient's daily life. Patients can wear smartwatches, rings, or headsets most of the time. Wearable devices using *ECG* data were used to evaluate PD in another 6 studies in a review [21]. Among these, results on statistical significance were inconsistent. However, some studies included Holter monitors as wearable devices, and they were not set in the patient's living environment, nor did they make use of the internet. In these studies, researchers found that heart rate variability (HRV) [22] can demonstrate the association between cardiac autonomic dysregulation and PD. Another survey by Cruz et al [23] used wearable and mobile systems to evaluate the severity of PA symptoms in correlation with physiological parameters. These parameters included the heart rate (HR), breathing rate, HRV, core temperature, and activities. However, it did not yield a model to predict PAs. The effect size and testing duration were both limited.

Jacobson et al [24] used a multilayered ensemble deep learning model paired with wearable *actigraph* units to passively sense data to predict deterioration in anxiety disorder symptoms. The result showed a balanced accuracy of 68.7% and an area under the curve of 69.6%. However, this study aimed to predict the long-term anxiety prognosis of PD rather than PAs. In addition, we could not correlate its time-sequence anxiety level with actigraphy. Sakamoto et al [25] used watches to detect PAs in 16 patients for 2 weeks. They found positive correlations between the PA frequency, locomotor activity ( $r=0.55$ ), and Hamilton Anxiety Rating Scale (HAM-A) scores.

### Goal of This Study

The purpose of this study was to establish a real-time PA prediction model. Data sources included clinical scales, diagnostic information, wearable devices, and environmental factors. We also compared the prediction importance between different data sources.

## Methods

### Participants

We recruited 59 participants from the En Chu Kong Hospital, Taiwan, psychiatric clinics between June 2020 and April 2021. The inclusion criteria were (1) a primary diagnosis of PD by DSM-5, (2) age more than 20 years, and (3) a basic ability to navigate smartwatch and mobile phone apps. Civil law defines an age of 20 years as becoming an adult in Taiwan. Below this age, the study required additional ethical regulation and opinions from participants' legal guardians, making the process more complicated.

The exclusion criteria were (1) current substance abuse, (2) cardiopulmonary incapacity, (3) limited mental capacity or total mental incapacity, and (4) acute suicidal ideation. This study required sufficient mental capacity on the part of participants to cooperate by continuously wearing smartwatches, properly maintaining the smartwatches, and completing regular, valid online questionnaires. Limited mental capacity implies that the person has difficulty understanding, remembering, or using the information to make or communicate a decision. Our team evaluated the participants' mental capacity during the diagnostic interview (DI), Mini International Neuropsychiatric Interview (MINI), and the process of informed consent by certified psychiatrists and nurse practitioners. The information about acute suicidal ideation was obtained from DIs and responses to questions in MINI part A and the preassessment Beck Depression Inventory (BDI).

### Study Approval

This study was approved and monitored by the institutional review board (ECKIRB1090305) of En Chu Kong Hospital. The research team securely stored all data according to the agreement, and privacy was protected by the Graduate Institute of Biomedical Electronics and Bioinformatics at National Taiwan University, Taiwan.

### Data Collection

The data we collected contained physiological data, environmental data, and questionnaire data. We obtained physiological data from the wearable device, which captured the participants' steps, distance traveled, floors climbed, HR in different states, and time of different sleep stages. The HR states captured during the monitoring period included (1) the minimum HR values, (2) the maximum HR values, (3) the average HR during the past 7 days, and (4) the average HR at rest, all in beats per minute (bpm). The different stages of sleep captured included (1) deep, (2) light, (3) rapid eye movement (REM), and (4) awake stages, all in seconds.

We obtained environmental data from the Environmental Protection Administration's Environmental Open Data Platform. We located the nearest environmental monitoring station according to each participant's residential address. These data were the air quality index (AQI), SO<sub>2</sub> subindex, CO subindex, particulate matter 1.0 microns (PM<sub>1.0</sub>) subindex, NO<sub>2</sub> subindex, and particulate matter 2.5 microns (PM<sub>2.5</sub>) subindex. We

collected these data every day to map the data from the smartwatches.

The questionnaire involved the Panic Disorder Severity Scale (PDSS), BDI, BAI, STAI, and MINI. Psychiatric professionals use MINI to screen the participants for mental comorbidities at the first DI.

### PDSS-SR, Chinese Version

Houck et al [26] developed the PDSS Self-Report version (PDSS-SR) in 2002, with the Chinese version [27] validated in 2020. This assessment includes 7 items: PA frequency, distress, anticipatory anxiety, agoraphobic fear, avoidance of panic-related bodily sensations, work impairment, and social impairment. Based on their rating on a 5-point scale, 0 indicated "not at all" and 1-4 indicated "mild," "moderate," "severe," and "extreme," respectively [27]. The first question in the PDSS-SR is, "How many panic and limited-symptom attacks did you have during the week?"

The prediction model ground truth (labeling) was "True" if the first question to the PDSS-SR was answered with 1, 2, 3, or 4 and "False" if the answer was 0. We sought to detect whether participants had experienced any PAs in the previous week. The PDSS-SR was collected at 2-week intervals for 1 continuous year via a mobile app or over the phone.

### BDI and BAI

The BDI II [28] measures the severity of depression using 21 questions. Each question has 4 choices (0, 1, 2, and 3): a higher score represents a more depressing description. The cut-off points of the sum are *minimal* (0-13), *mild* (14-19), *moderate* (20-28), and *severe* (29-63) depressive symptoms. The BAI II [29] measures the severity of anxiety using 21 questions. Each question has 4 choices: 0, *not at all*; 1 *mildly, but it didn't bother me much*; 2, *moderately—it wasn't pleasant at times*; and 3, *severely—it bothered me a lot*. The cut-off points of the sum are *minimal* (0-7), *mild* (8-15), *moderate* (16-26), and *severe* (26-63) depressive symptoms.

### STAI-S and STAI-T

The STAI Chinese version [30,31] measures anxiety levels. The STAI differentiates the temporary condition of state anxiety (STAI-S) and the more general and long-standing quality of trait anxiety (STAI-T). The essential attributes evaluated by the STAI-S scale are feelings of tension, nervousness, and worry [30]. The 4-point STAI-S scale is as follows: 1, *not at all*; 2, *somewhat*; 3, *moderately so*; and 4, *very much so*. The 4-point STAI-T scale is as follows: 1, *rarely*; 2, *sometimes*; 3, *often*; and 4, *almost constantly*. The cut-off point is 41 for the STAI-S and 43 for the STAI-T for clinically significant anxiety state/trait symptoms.

Participants self-reported their STAI-S and STAI-T initially and every 2 weeks via a mobile app.

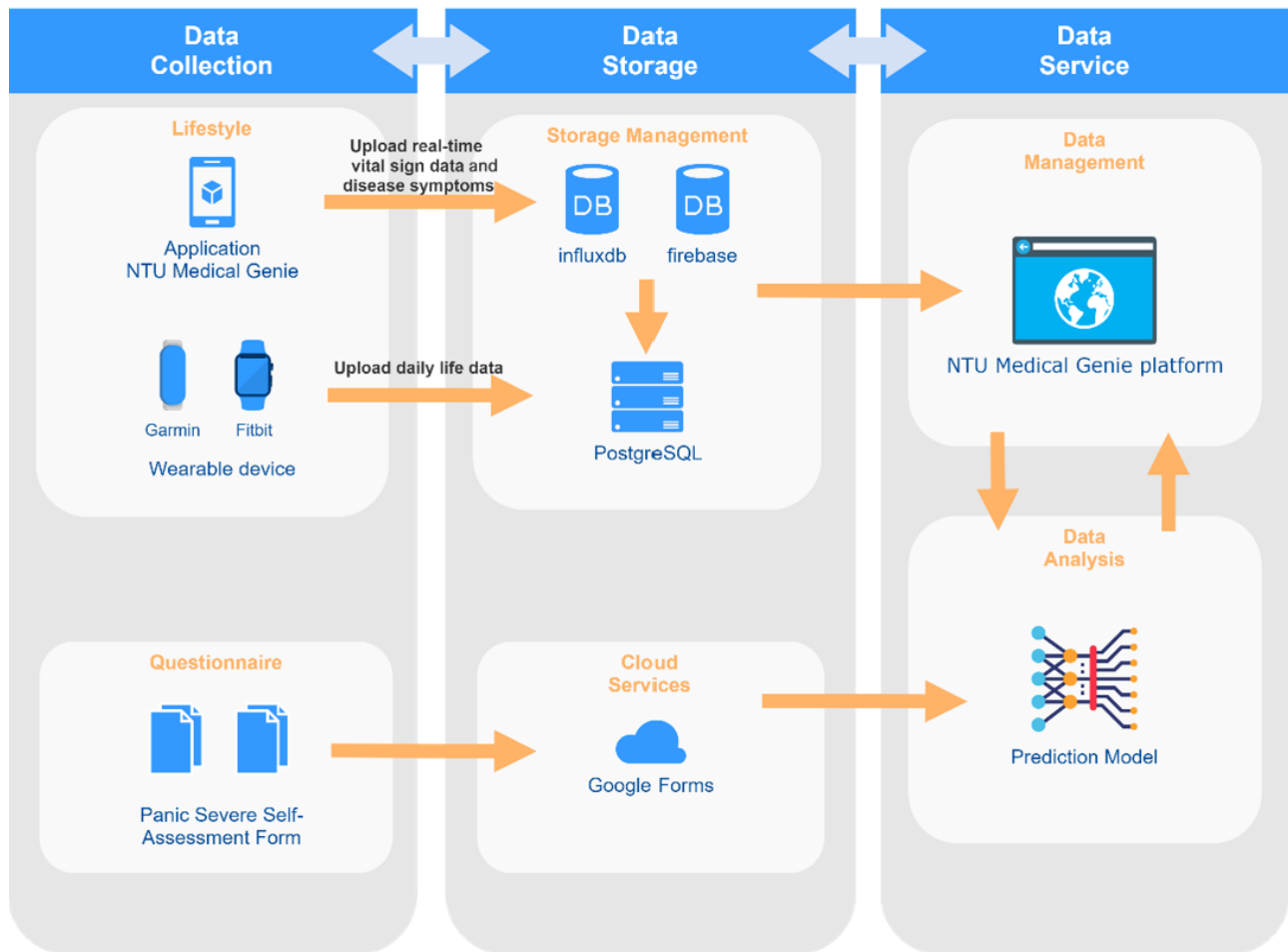
### System Architecture

The PA prediction system architecture contained 3 parts: data collection, data storage, and data service, as shown in Figure 1. For data collection, we included lifestyle data (physiological data) and questionnaire data. The wearable device (Garmin

Vivosmart 4) automatically collected the physiological information via Bluetooth and uploaded daily life data. In addition, we developed a smartphone app to collect real-time physiological data. Our team stored the daily life data in Postgres Structured Query Language and real-time physiological data in an influx database. Questionnaire data were collected via a Google form and stored in Google Drive.

We used the NTU Medical Genie platform for data service, management, and checking of participants' physiological data. Visualized data were also available on this platform, which helped the case manager to efficiently observe data. In addition, our team implemented the prediction model on the forum.

Figure 1. System architecture.

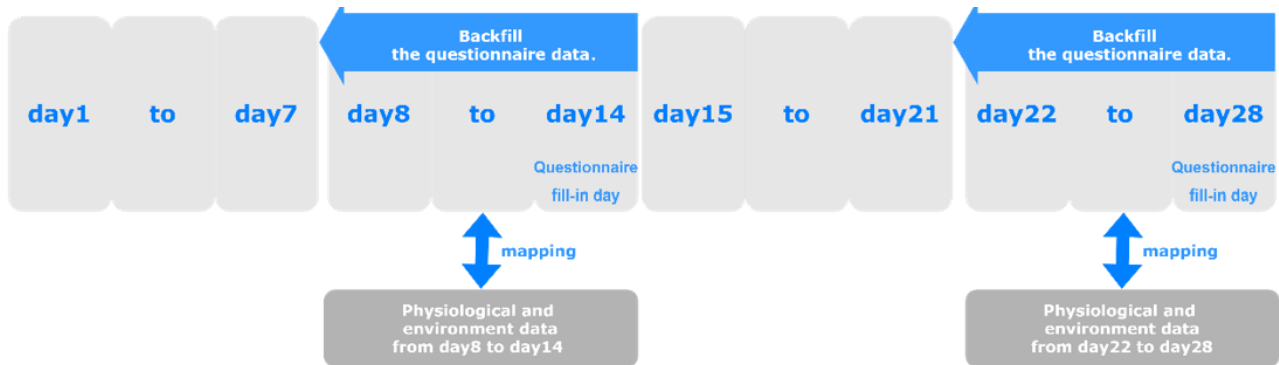


**Data Processing**

The data set we used to train the model was a combination of physiological data, environmental data, and questionnaire data. First, for missing values in the questionnaire data, we filled in

the average value of each question for each participant. Second, Figure 2 illustrates how we mapped physiological and questionnaire data. Participants filled out the questionnaire every 2 weeks.

Figure 2. Data mapping process.



We used backward filling to go back 7 days based on the questionnaire-filling date, after which we combined the questionnaire data with the physiological and environmental data. Participants responded to questionnaires based on their status in the past 7 days. The label summarizes "PA or not"; other situations reported in the questionnaire happened shortly before labeling time. In addition, 1 day corresponded to 1 data point, each of which serves as an individual observation for model training. If the label was true, we set each data point as true for 7 days and vice versa.

We experimented with several methods to mitigate the recall bias from questionnaires: First, the research assistants and clinicians followed up with the participants every 2 weeks over the telephone or through face-to-face interviews to ensure that the content of the questionnaires was consistent with the actual status in the previous week. Second, we examined the electronic medical records (EMRs) to determine whether the self-reported

content (PA or not) was consistent with medical notes. The study duration was 1 year; thus, the backfilling method allowed participants to report their mood every 2 weeks rather than that in a more intense time frame—once a week or daily—to facilitate their acceptance and adherence to the study.

After mapping all the data, we removed all data points for which physiological or environmental data were missing. This resulted in 3249 data points from June 2020 to March 15, 2021, as the training set and 974 data points from March 16 to April 2021 as the test set. We set the training and test sets in different time frames because it is closer to the clinical scenario. We aim to deploy this model in the future to mixed samples, both familiar and new patients. With this arrangement, the test set performance would benefit from within-patient correlation and also patients unseen by the model.

In [Textbox 1](#), we present the final set of features used in the model.

**Textbox 1.** Final set of features used in the study model.

<p><b>Environmental factors</b></p> <ul style="list-style-type: none"> <li>• Air quality index (AQI)</li> <li>• SO<sub>2</sub> subindex</li> <li>• CO subindex</li> <li>• Particulate matter 1.0 microns (PM<sub>1,0</sub>) subindex</li> <li>• NO<sub>2</sub> subindex</li> <li>• Particulate matter 2.5 microns (PM<sub>2,5</sub>) subindex</li> </ul> <p><b>Physiological factors</b></p> <ul style="list-style-type: none"> <li>• Steps</li> <li>• Distance</li> <li>• Floors</li> <li>• Minimum heart rate (HR)</li> <li>• Maximum HR</li> <li>• Average HR</li> <li>• Resting HR</li> <li>• Sleep duration</li> <li>• Deep sleep duration</li> <li>• Light sleep duration</li> <li>• Rapid-eye-movement (REM) sleep duration</li> <li>• Awake duration</li> </ul> <p><b>Clinical questionnaires</b></p> <ul style="list-style-type: none"> <li>• Beck Depression Inventory (BDI)</li> <li>• Beck Anxiety Inventory (BAI)</li> <li>• State-Trait Anxiety Inventory (STAI); 40 answers</li> <li>• Panic Disorder Severity Scale (PDSS); 1 answer as the ground truth</li> <li>• Mini International Neuropsychiatric Interview (MINI)</li> </ul>
---

## Classification Models

To predict PAs, we experimented with machine learning classifiers, including random forests, decision trees, linear discriminant analysis (LDA), adaptive boosting (AdaBoost), extreme gradient boosting (XGBoost), and regularized greedy forests (RGFs). We implemented these models using Python 3.6.10 libraries and Scikit-learn 0.23.1. We used 10-fold cross-validation and grid search for optimization of modeling. In the random forest example, we initially used a grid search to

set up different combinations of hyperparameters. We tried “n\_estimators: [50, 100, 200, 300], min\_samples\_split: [1, 2, 5, 10], min\_samples\_leaf: [1, 2, 5, 10],” with a total of 64 possible combinations of hyperparameters. After 10-fold cross-validation, we randomly split 10 parts of an equal amount of data in the training set. Later, we used 9 parts as training and 1 as validation in a rotation. Eventually, we averaged the F1 score of these 10 validation results and chose the best hyperparameter combination. The results of this process are shown in [Table 1](#).

**Table 1.** Model hyperparameters.

Model	Hyperparameter	Value, n
Random forest	n_estimators	100
	min_samples_split	2
	min_samples_leaf	1
Decision tree	min_samples_split	2
	min_samples_leaf	1
LDA <sup>a</sup>	solver	lsqr
	shrinkage	auto
AdaBoost <sup>b</sup>	n_estimators	50
	learning_rate	1
XGBoost <sup>c</sup>	objective	binary:logistic
	learning_rate	0.0001
RGF <sup>d</sup>	max_leaf	1000
	algorithm	RGF_Sib
	test_interval	100

<sup>a</sup>LDA: linear discriminant analysis.

<sup>b</sup>AdaBoost: adaptive boosting.

<sup>c</sup>XGBoost: extreme gradient boosting.

<sup>d</sup>RGF: regularized greedy forest.

## Validation and Model Assessment

We used 20% of the training data to evaluate the model in terms of accuracy, sensitivity, specificity, and the F1 score. We also used the testing data set to assess the model's predictive ability with respect to data never seen by the training model.

We tried several percentages, and the split of 20% gave the highest accuracy of the training result. According to previous experience from machine learning, a 10%-30% range is ideal for optimization of modeling.

## Results

### Clinical Characteristics of Participants

[Table 2](#) summarizes participant demographic factors and comorbidities according to MINI and the initial clinical

questionnaires. Participant ages ranged from 20 to 74 years. The female-male ratio was 1.56. Nearly half (30/59, 51%) of the participants were comorbid with at least 1 psychiatric illness: agoraphobia (13/59, 22%) and general anxiety disorder (GAD; 19/59, 32%) were the 2 most common comorbidities. In addition, 4 (7%) of the 59 participants were comorbid with depression, and 4 (7%) were comorbid with posttraumatic stress disorder (PTSD). The initial mean range of the PDSS-SR was 8.2 (SD 5.3), indicating clinically significant PA symptoms. The initial mean BAI was 20.5 (SD 12.4), and the mean BDI was 13.6 (SD 9.8), revealing a state of mild-to-moderate anxiety and minimal-to-mild depression. The initial mean STAI-S score was 45.2 (SD 7.2), and the initial mean STAI-T score was 47.6 (SD 7.1). Both state and situational anxiety were clinically significant at the time of recruitment.

**Table 2.** Clinical characteristics of participants (N=59).

Characteristics	Value	Interpretation
<b>Age (years)</b>		
Mean (SD)	46.2 (14.7)	Participant ages ranged from 20 to 74 years.
Range	20.1-74.8	
<b>Gender, n (%)</b>		
Male	23 (39.0)	The female-to-male ratio was 1.56.
Female	36 (61.0)	
<b>Comorbidity, n (%)</b>		
Agoraphobia	13 (22.0)	Nearly half (n=30, 51%) of the participants were comorbid with at least 1 psychiatric illness. Agoraphobia (n=13, 22%) and GAD (n=19, 32%) were the 2 most common comorbidities.
GAD <sup>a</sup>	19 (32.2)	
Social anxiety disorder (SAD)	1 (1.7)	
Major depressive disorder (MDD)	4 (6.8)	
Bipolar disorder	1 (1.7)	
PTSD <sup>b</sup>	4 (6.8)	
Obsessive-compulsive disorder (OCD)	2 (3.4)	
Others <sup>c</sup>	2 (3.4)	
No comorbidity	29 (49.2)	
<b>Initial PDSS-SR<sup>d</sup></b>		
Mean (SD)	8.2 (5.3)	Clinically significant panic symptoms.
Range	0-23	
<b>Initial BDI<sup>e</sup></b>		
Mean (SD)	13.6 (9.8)	Minimal-to-mild depression.
Range	0-46	
<b>Initial BAI<sup>f</sup></b>		
Mean (SD)	20.5 (12.4)	Mild-to-moderate anxiety.
Range	1-44	
<b>Initial STAI-S<sup>g</sup></b>		
Mean (SD)	45.2 (7.2)	Clinically significant situational anxiety.
Range	33-69	
<b>Initial STAI-T<sup>h</sup></b>		
Mean (SD)	47.6 (7.1)	Clinically significant trait anxiety.
Range	32-65	

<sup>a</sup>GAD: general anxiety disorder.

<sup>b</sup>PTSD: posttraumatic stress disorder.

<sup>c</sup>Others: history of heroin use disorder (n=1, 1.7%), psychotic disorder (n=1, 1.7%).

<sup>d</sup>PDSS-SR: Panic Disorder Severity Scale Self-Report (>4 shows significant PD symptoms).

<sup>e</sup>BDI: Beck Anxiety Inventory (minimal, 0-13; mild, 14-19; moderate, 20-28; severe, 29-63).

<sup>f</sup>BAI: Beck Anxiety Inventory (minimal, 0-7; mild, 8-15; moderate, 16-25; severe, 26-63).

<sup>g</sup>STAI-S: State-Trait Anxiety Inventory state anxiety (scoring 20-80, >41 shows situational anxiety).

<sup>h</sup>STAI-T: State-Trait Anxiety Inventory trait anxiety (scoring 20-80, >44 shows trait anxiety).

### PDSS-SR Result

Of all 3249 data points in the training set, 2109 (64.91%) showed no PA (PDSS-SR Q1=0), 832 (25.61%) showed only mild-intensity PAs (PDSS-SR Q1=1), 231 (7.11%) showed moderate PAs (PDSS-SR Q1=2), 58 (1.79%) showed severe PAs (PDSS-SR Q1=3), and 52 (1.6%) showed extreme PAs (PDSS-SR Q1=4). In addition, 32 (68%) of 47 participants experienced at least 1 PA or limited symptoms, and 15 (32%) of 47 participants denied any PA or had limited symptoms. Of all 974 data points in the test set, 641 (65.8%) showed no PA, 267 (27.4%) showed mild PAs, 65 (6.7%) showed moderate PAs, 1 (0.1%) showed severe PAs, and none showed extreme PAs. In addition, 28 (54%) of 52 participants experienced at least 1 PA or limited symptoms, and 24 (46%) of 52 participants denied any PA symptoms. All participants received current low-dose escitalopram or sertraline as the main PD regimen. The ratio of PA and non-PA was similar in the training set (35.1% vs 64.9%) and the test set (34.2% vs 65.8%).

### Panic Attack Prediction Model

We initially used data from the training phase to evaluate model performance, and the accuracy and F1 score of the implemented training set were as follows: random forest (0.975 and 0.968, respectively), decision tree (0.949 and 0.936, respectively), LDA (0.746 and 0.647, respectively), AdaBoost (0.838 and 0.792, respectively), XGBoost (0.702 and 0.458, respectively), RGF (0.945 and 0.928, respectively). Table 3 presents the test set performance. The random forest offered the highest accuracy compared to other models, whether in training or in testing models. The area under the receiver operating characteristic (AUROC) curve of each prediction algorithm is shown in Figure 3.

We also tested the model with different combinations of data sets, as shown in Table 4. These results show that the prediction performance of the all-feature model is better than that of the physiological-environment model or the questionnaire model alone.

**Table 3.** Test set performance of each model with all features.

Model	Accuracy	AUROC <sup>a</sup>	Specificity	Sensitivity	Precision	F1 score
Random forest	0.813	0.871	0.938	0.574	0.827	0.677
Decision tree	0.705	0.674	0.772	0.577	0.568	0.572
LDA <sup>b</sup>	0.722	0.720	0.850	0.474	0.622	0.538
AdaBoost <sup>c</sup>	0.746	0.794	0.872	0.505	0.672	0.576
XGBoost <sup>d</sup>	0.674	0.763	0.913	0.213	0.559	0.309
RGF <sup>e</sup>	0.800	0.863	0.920	0.568	0.788	0.660

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

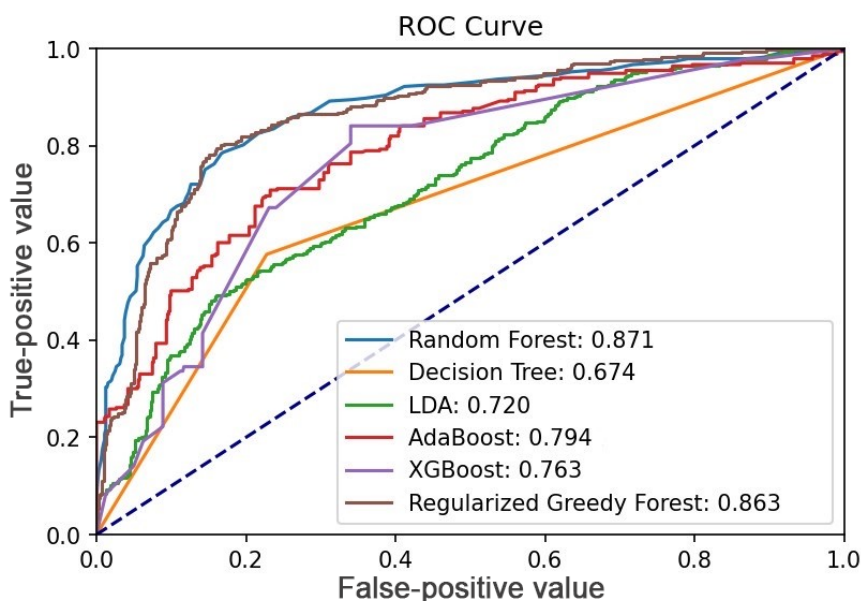
<sup>b</sup>LDA: linear discriminant analysis.

<sup>c</sup>AdaBoost: adaptive boosting.

<sup>d</sup>XGBoost: extreme gradient boosting.

<sup>e</sup>RGF: regularized greedy forest.

**Figure 3.** ROC curve analysis of prediction algorithms of test set. LDA: linear discriminant analysis; ROC: receiver operating characteristic.





**Table 4.** Test set performance of each model with various data set combinations.

Feature	Model	Accuracy	AUROC <sup>a</sup>	Specificity	Sensitivity	Precision	F1 score
All features	Random forest	0.813	0.872	0.938	0.574	0.827	0.677
Lifestyle and environment	RGF <sup>b</sup>	0.674	0.687	0.773	0.477	0.513	0.495
Questionnaire	RGF	0.771	0.843	0.858	0.617	0.712	0.661

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

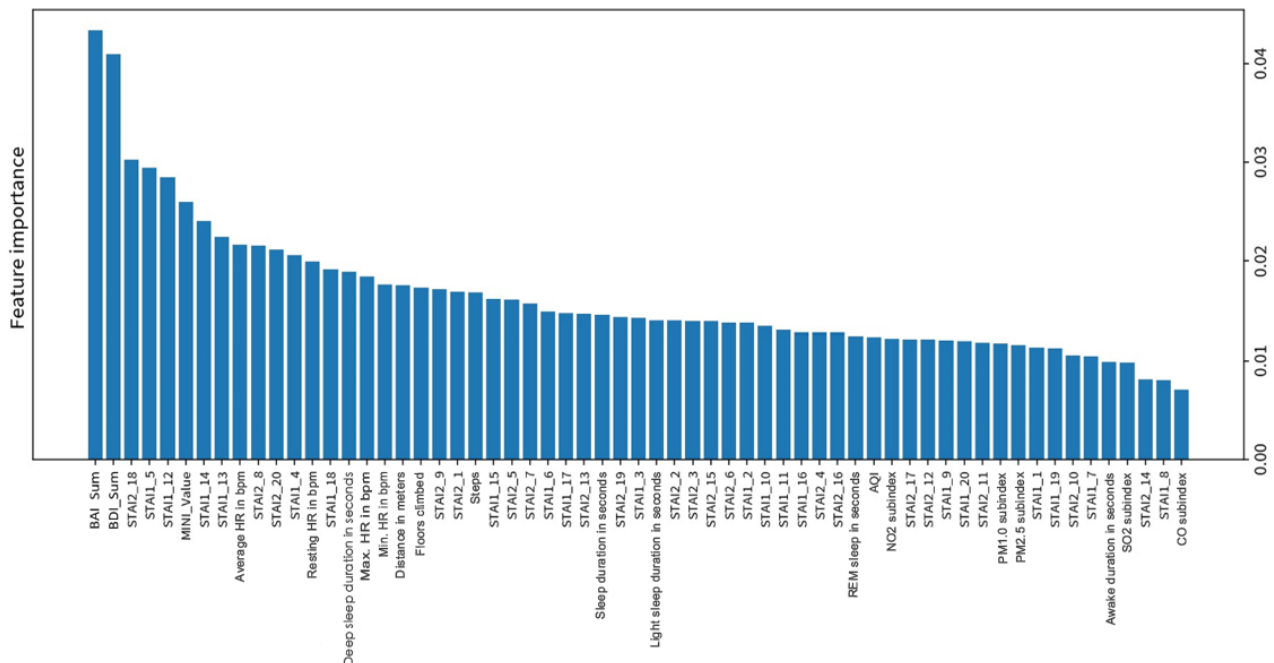
<sup>b</sup>RGF: regularized greedy forest.

## Feature Importance

Feature importance refers to a feature's importance level in model prediction: the larger the number, the more critical the

feature. Figure 4 shows the feature importance of the all-feature model. Questionnaire and physiological features, such as the BAI, BDI, STAI, MINI, average HR, resting HR, and deep sleep duration, were more critical than others in this prediction model.

**Figure 4.** Feature importance of the all-feature model from a random forest. AQI: air quality index; BAI: Beck Anxiety Inventory; BDI: Beck Depression Inventory; bpm: beats per minute; MINI: Mini International Neuropsychiatric Interview; PM<sub>1.0</sub>: particulate matter 1.0 microns; PM<sub>2.5</sub>: particulate matter 2.5 microns; REM: rapid eye movement; STAI: State-Trait Anxiety Inventory.



## Discussion

### Principal Findings

PAs can be predicted 1 week before occurrence by machine learning through clinical questionnaires, physiological data, and environmental data. Random forests yielded the best prediction accuracy (81.3%) on the test set. Overall, the test set accuracy was 67.4%-81.3% for various machine learning algorithms. The feature importance ranking from high to low was clinical questionnaires, physiological data, and environmental data in the training set. The essential features for PA prediction were the BDI, BAI, STAI, MINI, HR in different states, and deep sleep duration. The prediction performance of the all-feature model was better than that of the physiological-environment model or the questionnaire model alone. This also highlighted that wearable devices detecting HR or deep sleep duration could be a potential tool to predict PAs.

### Study Strengths

To the best of our knowledge, this is the first PA prediction model study evaluated in *real life* with a full year of *continuous monitoring*. We also provided *multifactor features* for PA prediction, including physiological factors via smartwatches, clinical questionnaires, and environmental factors. We collected the questionnaire data via an *internet-based* mobile app, which is more accessible for most participants. Most participants gave positive feedback after learning to self-monitor their emotional and physiological states through wearables and regular questionnaires under supervision.

### Study Limitations

First, the sample size (N=59) was limited because this study required participants' intensive cooperation. However, at the time of this study, 59 was a relatively large number in the known literature on using wearables for PA prediction [23,25]. Second,

the prediction model was derived primarily from participants under regular medication in a single hospital. The performance would benefit from within-patient correlation; however, more external testing is needed for those patients unseen by the model. Third, the PA ground-truth labels were from the PDSS-SR questionnaires [32]. The participants' memory recall could be biased while tracing back to previous events; labeling validity also depends on the participants' understanding of the nature of PAs. To minimize these problems, we provided comprehensive psychoeducation to participants before this trial. Research teams used telephone follow-ups every 2 weeks to determine whether there were obvious outliers or missing data due to technical problems with the participants. Finally, according to the current study design, the PA prediction result applies only to patients with an established diagnosis of PD.

### Comparison With Prior Work

This study used multifactorial variables. Compared to previous PA or PD studies [13,33], our study combined questionnaire data with physiological and environmental data, resulting in superior prediction results as compared to a single data source (see Table 4). Prior work [21] focused on PA prediction was in clinician-monitored environments. However, wearables, such as smartwatches, and mobile apps [34,35] can be used in real-life situations, increasing ecological validity. In previous studies, the wearables' testing duration was often days to weeks or cross-sectional [23], detecting few real-time PA events. Our study continued for 1 year and detected PA events in 1140 (35.09%) of all 3249 data points, a more balanced data distribution, making machine learning a possible tool for prediction.

In our experience, regular online questionnaires require intensive cooperation from participants and supervision by clinicians, which may be burdensome [36]. Wearable devices, however, are easier for autorecording with a real-time notification function. The use of combined methods for PA prediction needs further feasibility studies in actual clinical settings. Several studies have correlated the HRV to trait anxiety and depressiveness [37,38]. Thus, it is possible to merge the measurement of trait anxiety (STAI-T) and depressiveness (BDI) from questionnaires into wearables with an HRV-detecting function to provide information for prediction.

Our team also found that the AQI is less critical than questionnaires and wearable sensor data, which differs from the result, showing a significant relationship between air pollution

and PAs in emergency visits [6]. The difference needs further evaluation because the nearest environmental monitoring station to the residential address may not reflect the actual location where each participant stayed. Using the Global Positioning System or air quality sensors located at individual participants' homes is one way to address this problem.

### Clinical Suggestions

To better predict PAs, it is possible to use multifactorial items from clinical questionnaires and physiological and environmental data. Among these, clinical questionnaires are more crucial than their physiological-environmental counterparts. It is also beneficial to collect information from baseline anxiety and depression, trait anxiety, the number of comorbid psychiatric diagnoses, the average and resting HR, and deep sleep duration as a reference to predict recurrent PAs for patients with PD.

### Future Work

First, we will collect more participants to increase the effect size and sample heterogeneity. Currently, we do not clearly understand the correlation between PA symptoms and individual features. We suggest using an explainable model and combining questionnaires with real-time HRV data to establish a model to predict PAs hours before their occurrence.

### Conclusion

This prospective study introduced a 7-day prediction model for PAs based on machine learning using wearable devices, online questionnaires, and environmental data for a combinational assessment of PD, continuously monitoring samples from real-life settings for 1 year.

It is possible to predict PAs 7 days before the fact by using a combination of all data from questionnaires, physiological data, and environmental data. The prediction accuracy was 67.4%-81.3% for the test set from various machine learning algorithms, among which random forests offered the highest accuracy compared to other models. The prediction performance of the all-feature model is better than the physiological-environment model or questionnaire model alone. The features that contributed most to the prediction models are the BAI, BDI, STAI, MINI, average HR, resting HR, and deep sleep duration. However, current findings apply only to patients with an established diagnosis of PD. More external testing is also needed.

---

### Acknowledgments

The Ministry of Science and Technology of Taiwan supported this research (grant MOST 110-2634-F-002-032-) and English editing (108-2926-I-002-002-MY4).

We thank our colleagues in the Department of Psychiatry of En Chu Kong Hospital and the Graduate Institute of Biomedical Electronics and Bioinformatics of National Taiwan University for their assistance with data collection and analysis.

---

## Authors' Contributions

C-HT designed this study and was responsible for participant recruitment and paper processing; P-CC and D-SL were responsible for model training and completed parts of the method and results; C-TW, T-TH, and D-LC designed the platform, wearables, and mobile app system; Y-YK was responsible for the enrollment, contact, and participant follow-ups; and FL directed the project.

## Conflicts of Interest

The authors declare no potential conflicts of interest in this paper's research, authorship, or publication.

## References

1. de Jonge P, Roest AM, Lim CCW, Florescu SE, Bromet EJ, Stein DJ, et al. Cross-national epidemiology of panic disorder and panic attacks in the world mental health surveys. *Depress Anxiety* 2016 Dec;33(12):1155-1177 [FREE Full text] [doi: [10.1002/da.22572](https://doi.org/10.1002/da.22572)] [Medline: [27775828](https://pubmed.ncbi.nlm.nih.gov/27775828/)]
2. Chang H, Pan C, Chen P, Chen Y, Su S, Tsai S, et al. Identification and medical utilization of newly diagnosed panic disorder: a nationwide case-control study. *J Psychosom Res* 2019 Oct;125:109815. [doi: [10.1016/j.jpsychores.2019.109815](https://doi.org/10.1016/j.jpsychores.2019.109815)] [Medline: [31450125](https://pubmed.ncbi.nlm.nih.gov/31450125/)]
3. Chen M, Tsai S. Treatment-resistant panic disorder: clinical significance, concept and management. *Prog Neuropsychopharmacol Biol Psychiatry* 2016 Oct 03;70:219-226 [FREE Full text] [doi: [10.1016/j.pnpbp.2016.02.001](https://doi.org/10.1016/j.pnpbp.2016.02.001)] [Medline: [26850787](https://pubmed.ncbi.nlm.nih.gov/26850787/)]
4. American Psychiatric Association. *Diagnostic and Statistical Manual of Mental Disorders*, 5th edition. Washington, DC: American Psychiatric Association; 2013.
5. Potter CM, Wong J, Heimberg RG, Blanco C, Liu S, Wang S, et al. Situational panic attacks in social anxiety disorder. *J Affect Disord* 2014;167:1-7 [FREE Full text] [doi: [10.1016/j.jad.2014.05.044](https://doi.org/10.1016/j.jad.2014.05.044)] [Medline: [25082106](https://pubmed.ncbi.nlm.nih.gov/25082106/)]
6. Cho J, Choi YJ, Sohn J, Suh M, Cho S, Ha KH, et al. Ambient ozone concentration and emergency department visits for panic attacks. *J Psychiatr Res* 2015 Mar;62:130-135 [FREE Full text] [doi: [10.1016/j.jpsychores.2015.01.010](https://doi.org/10.1016/j.jpsychores.2015.01.010)] [Medline: [25669697](https://pubmed.ncbi.nlm.nih.gov/25669697/)]
7. Trushna T, Dhiman V, Raj D, Tiwari R. Effects of ambient air pollution on psychological stress and anxiety disorder: a systematic review and meta-analysis of epidemiological evidence. *Rev Environ Health* 2021 Dec 20;36(4):501-521 [FREE Full text] [doi: [10.1515/reveh-2020-0125](https://doi.org/10.1515/reveh-2020-0125)] [Medline: [34821119](https://pubmed.ncbi.nlm.nih.gov/34821119/)]
8. Iurato S, Carrillo-Roa T, Arloth J, Czamara D, Diener-Hölzl L, Lange J, et al. DNA methylation signatures in panic disorder. *Transl Psychiatry* 2017 Dec 18;7(12):1287 [FREE Full text] [doi: [10.1038/s41398-017-0026-1](https://doi.org/10.1038/s41398-017-0026-1)] [Medline: [29249830](https://pubmed.ncbi.nlm.nih.gov/29249830/)]
9. Robinaugh DJ, Ward MJ, Toner ER, Brown ML, Losiewicz OM, Bui E, et al. Assessing vulnerability to panic: a systematic review of psychological and physiological responses to biological challenges as prospective predictors of panic attacks and panic disorder. *Gen Psych* 2019 Dec 23;32(6):e100140-e101463 [FREE Full text] [doi: [10.1136/gpsych-2019-100140](https://doi.org/10.1136/gpsych-2019-100140)] [Medline: [3192208](https://pubmed.ncbi.nlm.nih.gov/3192208/)]
10. Norton G, Pidlubny S, Norton PJ. Prediction of panic attacks and related variables. *Behav Therapy* 1999;30(2):319-330 [FREE Full text] [doi: [10.1016/S0005-7894\(99\)80011-2](https://doi.org/10.1016/S0005-7894(99)80011-2)]
11. Schmidt NB, Zvolensky MJ, Maner JK. Anxiety sensitivity: prospective prediction of panic attacks and axis I pathology. *J Psychiatr Res* 2006 Dec;40(8):691-699 [FREE Full text] [doi: [10.1016/j.jpsychores.2006.07.009](https://doi.org/10.1016/j.jpsychores.2006.07.009)] [Medline: [16956622](https://pubmed.ncbi.nlm.nih.gov/16956622/)]
12. Jacobs G, Latham L, Brown M. Test-retest reliability of the state-trait personality inventory and the anger expression scale. *Anxiety Res* 1988 Jan;1(3):263-265 [FREE Full text] [doi: [10.1080/08917778808248724](https://doi.org/10.1080/08917778808248724)]
13. Liu Y, Sareen J, Bolton J, Wang J. Development and validation of a risk-prediction algorithm for the recurrence of panic disorder. *Depress Anxiety* 2015 May;32(5):341-348 [FREE Full text] [doi: [10.1002/da.22359](https://doi.org/10.1002/da.22359)] [Medline: [25774487](https://pubmed.ncbi.nlm.nih.gov/25774487/)]
14. Lueken U, Straube B, Yang Y, Hahn T, Beesdo-Baum K, Wittchen H, et al. Separating depressive comorbidity from panic disorder: a combined functional magnetic resonance imaging and machine learning approach. *J Affect Disord* 2015 Sep 15;184:182-192 [FREE Full text] [doi: [10.1016/j.jad.2015.05.052](https://doi.org/10.1016/j.jad.2015.05.052)] [Medline: [26093832](https://pubmed.ncbi.nlm.nih.gov/26093832/)]
15. Hahn T, Kircher T, Straube B, Wittchen HU, Konrad C, Ströhle A, et al. Predicting treatment response to cognitive behavioral therapy in panic disorder with agoraphobia by integrating local neural information. *JAMA Psychiatry* 2015 Jan 15;72(1):68-74 [FREE Full text] [doi: [10.1001/jamapsychiatry.2014.1741](https://doi.org/10.1001/jamapsychiatry.2014.1741)] [Medline: [25409415](https://pubmed.ncbi.nlm.nih.gov/25409415/)]
16. Grambal A, Tüdös Z, Kamaradová D, Hlušík P, Praško J. Treatment effect prediction in panic disorder patients: fMRI study. *Eur Psychiatry* 2013 Jan;28:1 [FREE Full text] [doi: [10.1016/S0924-9338\(13\)76550-7](https://doi.org/10.1016/S0924-9338(13)76550-7)]
17. de Carvalho MR, Dias GP, Cosci F, de-Melo-Neto VL, Bevilacqua MCDN, Gardino PF, et al. Current findings of fMRI in panic disorder: contributions for the fear neurocircuitry and CBT effects. *Expert Rev Neurother* 2010 Feb;10(2):291-303 [FREE Full text] [doi: [10.1586/ern.09.161](https://doi.org/10.1586/ern.09.161)] [Medline: [20136384](https://pubmed.ncbi.nlm.nih.gov/20136384/)]
18. Sundermann B, Bode J, Lueken U, Westphal D, Gerlach AL, Straube B, et al. Support vector machine analysis of functional magnetic resonance imaging of interoception does not reliably predict individual outcomes of cognitive behavioral therapy in panic disorder with agoraphobia. *Front Psychiatry* 2017;8:99 [FREE Full text] [doi: [10.3389/fpsy.2017.00099](https://doi.org/10.3389/fpsy.2017.00099)] [Medline: [28649205](https://pubmed.ncbi.nlm.nih.gov/28649205/)]

19. Hayashi K, Makino M, Hashizume M, Nakano K, Tsuboi K. Electroencephalogram abnormalities in panic disorder patients: a study of symptom characteristics and pathology. *Biopsychosoc Med* 2010 Aug 23;4:9 [FREE Full text] [doi: [10.1186/1751-0759-4-9](https://doi.org/10.1186/1751-0759-4-9)] [Medline: [20731860](https://pubmed.ncbi.nlm.nih.gov/20731860/)]
20. Carvalho MRD, Velasques BB, Cagy M, Marques JB, Teixeira S, Nardi AE, et al. Electroencephalographic findings in panic disorder. *Trends Psychiatry Psychother* 2013 Dec;35(4):238-251 [FREE Full text] [doi: [10.1590/2237-6089-2013-0012](https://doi.org/10.1590/2237-6089-2013-0012)] [Medline: [27000470](https://pubmed.ncbi.nlm.nih.gov/27000470/)]
21. Elgendi M, Menon C. Assessing anxiety disorders using wearable devices: challenges and future directions. *Brain Sci* 2019 Mar 01;9(3):50 [FREE Full text] [doi: [10.3390/brainsci9030050](https://doi.org/10.3390/brainsci9030050)] [Medline: [30823690](https://pubmed.ncbi.nlm.nih.gov/30823690/)]
22. Chang H, Chang C, Tzeng N, Kuo TBJ, Lu R, Huang S. Decreased cardiac vagal control in drug-naive patients with panic disorder: a case-control study in Taiwan. *Asia Pac Psychiatry* 2013 Jun;5(2):80-89 [FREE Full text] [doi: [10.1111/appy.12032](https://doi.org/10.1111/appy.12032)] [Medline: [23857808](https://pubmed.ncbi.nlm.nih.gov/23857808/)]
23. Cruz L, Rubin J, Abreu R, Ahern S, Eldardiry H, Bobrow D. A wearable and mobile intervention delivery system for individuals with panic disorder. 2015 Presented at: Proceedings of the 14th International Conference on Mobile and Ubiquitous Multimedia; 2015; Linz, Austria p. 175-182. [doi: [10.1145/2836041.2836058](https://doi.org/10.1145/2836041.2836058)]
24. Jacobson NC, Lekkas D, Huang R, Thomas N. Deep learning paired with wearable passive sensing data predicts deterioration in anxiety disorder symptoms across 17-18 years. *J Affect Disord* 2021 Mar 01;282:104-111 [FREE Full text] [doi: [10.1016/j.jad.2020.12.086](https://doi.org/10.1016/j.jad.2020.12.086)] [Medline: [33401123](https://pubmed.ncbi.nlm.nih.gov/33401123/)]
25. Sakamoto N, Yoshiuchi K, Kikuchi H, Takimoto Y, Kaiya H, Kumano H, et al. Panic disorder and locomotor activity. *Biopsychosoc Med* 2008 Nov 18;2:23 [FREE Full text] [doi: [10.1186/1751-0759-2-23](https://doi.org/10.1186/1751-0759-2-23)] [Medline: [19017383](https://pubmed.ncbi.nlm.nih.gov/19017383/)]
26. Houck PR, Spiegel DA, Shear MK, Rucci P. Reliability of the self-report version of the panic disorder severity scale. *Depress Anxiety* 2002;15(4):183-185. [doi: [10.1002/da.10049](https://doi.org/10.1002/da.10049)] [Medline: [12112724](https://pubmed.ncbi.nlm.nih.gov/12112724/)]
27. Liu X, Xu T, Chen D, Yang C, Wang P, Huang X, et al. Reliability, validity and cut-off score of the Chinese version of the panic disorder severity scale self-report form in patients with panic disorder. *BMC Psychiatry* 2020 Apr 15;20(1):170 [FREE Full text] [doi: [10.1186/s12888-020-02560-w](https://doi.org/10.1186/s12888-020-02560-w)] [Medline: [32295552](https://pubmed.ncbi.nlm.nih.gov/32295552/)]
28. Wang Z, Yuan CM, Huang J, Li ZZ, Chen J, Zhang HY. Reliability and validity of the Chinese version of Beck Depression Inventory-II among depression patients. *Chin Ment Health J* 2011;25(6):476-480.
29. Cheng S, Wong C, Wong K, Chong G, Wong M, Chang S. A study of psychometric properties, normative scores, and factor structure of the Beck Anxiety Inventory: the Chinese version. *Chin J Clin Psychol* 2002;10(1):4-6.
30. Kuo-Chang W, Feng-Chiao C. An investigation of multidimensional factorial validity of the Chinese VERSION of State-Trait Anxiety Inventory. *Psychol Test* 2016;63(4):287-313.
31. Spielberger C. State-Trait Anxiety Inventory. In: *The Corsini Encyclopedia of Psychology*. Hoboken, NJ: John Wiley & Sons; 2010.
32. Alfonsson S, Maathz P, Hursti T. Interformat reliability of digital psychiatric self-report questionnaires: a systematic review. *J Med Internet Res* 2014 Dec 03;16(12):e268 [FREE Full text] [doi: [10.2196/jmir.3395](https://doi.org/10.2196/jmir.3395)] [Medline: [25472463](https://pubmed.ncbi.nlm.nih.gov/25472463/)]
33. Meuret AE, Rosenfield D, Wilhelm FH, Zhou E, Conrad A, Ritz T, et al. Do unexpected panic attacks occur spontaneously? *Biol Psychiatry* 2011 Nov 15;70(10):985-991 [FREE Full text] [doi: [10.1016/j.biopsych.2011.05.027](https://doi.org/10.1016/j.biopsych.2011.05.027)] [Medline: [21783179](https://pubmed.ncbi.nlm.nih.gov/21783179/)]
34. Bakker D, Kazantzis N, Rickwood D, Rickard N. Mental health smartphone apps: review and evidence-based recommendations for future developments. *JMIR Ment Health* 2016 Mar 01;3(1):e7 [FREE Full text] [doi: [10.2196/mental.4984](https://doi.org/10.2196/mental.4984)] [Medline: [26932350](https://pubmed.ncbi.nlm.nih.gov/26932350/)]
35. Beiwinkel T, Kindermann S, Maier A, Kerl C, Moock J, Barbian G, et al. Using smartphones to monitor bipolar disorder symptoms: a pilot study. *JMIR Ment Health* 2016 Jan 06;3(1):e2 [FREE Full text] [doi: [10.2196/mental.4560](https://doi.org/10.2196/mental.4560)] [Medline: [26740354](https://pubmed.ncbi.nlm.nih.gov/26740354/)]
36. Rolstad S, Adler J, Rydén A. Response burden and questionnaire length: is shorter better? A review and meta-analysis. *Value Health* 2011 Dec;14(8):1101-1108 [FREE Full text] [doi: [10.1016/j.jval.2011.06.003](https://doi.org/10.1016/j.jval.2011.06.003)] [Medline: [22152180](https://pubmed.ncbi.nlm.nih.gov/22152180/)]
37. Shinba T, Kariya N, Matsui Y, Ozawa N, Matsuda Y, Yamamoto K. Decrease in heart rate variability response to task is related to anxiety and depressiveness in normal subjects. *Psychiatry Clin Neurosci* 2008 Oct;62(5):603-609 [FREE Full text] [doi: [10.1111/j.1440-1819.2008.01855.x](https://doi.org/10.1111/j.1440-1819.2008.01855.x)] [Medline: [18950382](https://pubmed.ncbi.nlm.nih.gov/18950382/)]
38. Miu AC, Heilman RM, Miclea M. Reduced heart rate variability and vagal tone in anxiety: trait versus state, and the effects of autogenic training. *Auton Neurosci* 2009 Jan 28;145(1-2):99-103. [doi: [10.1016/j.autneu.2008.11.010](https://doi.org/10.1016/j.autneu.2008.11.010)] [Medline: [19059813](https://pubmed.ncbi.nlm.nih.gov/19059813/)]

## Abbreviations

- AdaBoost:** adaptive boosting
- AQI:** air quality index
- AUROC:** area under the receiver operating characteristic
- BAI:** Beck Anxiety Inventory
- BDI:** Beck Depression Inventory
- bpm:** beats per minute

**DI:** diagnostic interview  
**DSM-5:** *Diagnostic and Statistical Manual of Mental Disorders, 5th edition*  
**ECG:** electrocardiogram  
**EEG:** electroencephalogram  
**EMR:** electronic medical record  
**fMRI:** functional magnetic resonance imaging  
**HAM-A:** Hamilton Anxiety Rating Scale  
**HAM-D:** Hamilton Depression Rating Scale  
**HR:** heart rate  
**HRV:** heart rate variability  
**LDA:** linear discriminant analysis  
**MINI:** Mini International Neuropsychiatric Interview  
**PA:** panic attack  
**PD:** panic disorder  
**PDSS-SR:** Panic Disorder Severity Scale (PDSS), Self-Report  
**PM<sub>1.0</sub>:** particulate matter 1.0 microns  
**PM<sub>2.5</sub>:** particulate matter 2.5 microns  
**PTSD:** posttraumatic stress disorder  
**RGF:** regularized greedy forest  
**REM:** rapid eye movement  
**ROC:** receiver operating characteristic  
**STAI-S:** State-Trait Anxiety Inventory state anxiety  
**STAI-T:** State-Trait Anxiety Inventory trait anxiety  
**XGBoost:** extreme gradient boosting

*Edited by C Lovis; submitted 23.08.21; peer-reviewed by YH Sheu, K Mathiasen, A Joshi, JL Domínguez-Olmedo; comments to author 14.09.21; revised version received 08.11.21; accepted 02.01.22; published 15.02.22.*

*Please cite as:*

Tsai CH, Chen PC, Liu DS, Kuo YY, Hsieh TT, Chiang DL, Lai F, Wu CT

*Panic Attack Prediction Using Wearable Devices and Machine Learning: Development and Cohort Study*

*JMIR Med Inform 2022;10(2):e33063*

URL: <https://medinform.jmir.org/2022/2/e33063>

doi: [10.2196/33063](https://doi.org/10.2196/33063)

PMID: [35166679](https://pubmed.ncbi.nlm.nih.gov/35166679/)

©Chan-Hen Tsai, Pei-Chen Chen, Ding-Shan Liu, Ying-Ying Kuo, Tsung-Ting Hsieh, Dai-Lun Chiang, Feipei Lai, Chia-Tung Wu. Originally published in JMIR Medical Informatics (<https://medinform.jmir.org>), 15.02.2022. This is an open-access article distributed under the terms of the Creative Commons Attribution License (<https://creativecommons.org/licenses/by/4.0/>), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Medical Informatics, is properly cited. The complete bibliographic information, a link to the original publication on <https://medinform.jmir.org/>, as well as this copyright and license information must be included.

Original Paper

# Early Prediction of Functional Outcomes After Acute Ischemic Stroke Using Unstructured Clinical Text: Retrospective Cohort Study

Sheng-Feng Sung<sup>1,2</sup>, MD, PhD; Cheng-Yang Hsieh<sup>3</sup>, MD, PhD; Ya-Han Hu<sup>4</sup>, PhD

<sup>1</sup>Division of Neurology, Department of Internal Medicine, Ditmanson Medical Foundation Chia-Yi Christian Hospital, Chiayi City, Taiwan

<sup>2</sup>Department of Nursing, Min-Hwei Junior College of Health Care Management, Tainan, Taiwan

<sup>3</sup>Department of Neurology, Tainan Sin Lau Hospital, Tainan, Taiwan

<sup>4</sup>Department of Information Management, National Central University, Taoyuan City, Taiwan

**Corresponding Author:**

Ya-Han Hu, PhD

Department of Information Management

National Central University

No 300, Zhongda Rd

Zhongli District

Taoyuan City, 320317

Taiwan

Phone: 886 34227151 ext 66560

Email: [yhhu@mgt.ncu.edu.tw](mailto:yhhu@mgt.ncu.edu.tw)

## Abstract

**Background:** Several prognostic scores have been proposed to predict functional outcomes after an acute ischemic stroke (AIS). Most of these scores are based on structured information and have been used to develop prediction models via the logistic regression method. With the increased use of electronic health records and the progress in computational power, data-driven predictive modeling by using machine learning techniques is gaining popularity in clinical decision-making.

**Objective:** We aimed to investigate whether machine learning models created by using unstructured text could improve the prediction of functional outcomes at an early stage after AIS.

**Methods:** We identified all consecutive patients who were hospitalized for the first time for AIS from October 2007 to December 2019 by using a hospital stroke registry. The study population was randomly split into a training (n=2885) and test set (n=962). Free text in histories of present illness and computed tomography reports was transformed into input variables via natural language processing. Models were trained by using the extreme gradient boosting technique to predict a poor functional outcome at 90 days poststroke. Model performance on the test set was evaluated by using the area under the receiver operating characteristic curve (AUC).

**Results:** The AUCs of text-only models ranged from 0.768 to 0.807 and were comparable to that of the model using National Institutes of Health Stroke Scale (NIHSS) scores (0.811). Models using both patient age and text achieved AUCs of 0.823 and 0.825, which were similar to those of the model containing age and NIHSS scores (0.841); the model containing preadmission comorbidities, level of consciousness, age, and neurological deficit (PLAN) scores (0.837); and the model containing Acute Stroke Registry and Analysis of Lausanne (ASTRAL) scores (0.840). Adding variables from clinical text improved the predictive performance of the model containing age and NIHSS scores, the model containing PLAN scores, and the model containing ASTRAL scores (the AUC increased from 0.841 to 0.861, from 0.837 to 0.856, and from 0.840 to 0.860, respectively).

**Conclusions:** Unstructured clinical text can be used to improve the performance of existing models for predicting poststroke functional outcomes. However, considering the different terminologies that are used across health systems, each individual health system may consider using the proposed methods to develop and validate its own models.

(*JMIR Med Inform* 2022;10(2):e29806) doi:[10.2196/29806](https://doi.org/10.2196/29806)

**KEYWORDS**

acute ischemic stroke; bag-of-words; extreme gradient boosting; machine learning; MetaMap; natural language processing; outcome prediction; text classification; unstructured clinical text

## Introduction

Stroke is a common and serious neurologic disorder. Approximately 1 out of every 4 adults aged  $\geq 25$  years will experience a stroke in their lifetime [1]. Despite recent and emerging advances in the acute treatment of strokes, more than half of patients with stroke still experience an unfavorable outcome, which can result in permanent functional dependency or even death [2]. In clinical practice, having a handy and readily available prognostic tool is desirable for clinical decision-making and resource allocation. Prognostic understanding is of direct clinical relevance and is essential for informing goals-of-care discussions. It also facilitates discharge planning, communication, and postdischarge support.

Several prognostic scores have been developed to predict functional outcomes following an acute stroke. Most of these scores use similar input variables for their predictions. As functional outcomes are largely determined by age and stroke severity [3], these two variables are almost always included in existing prognostic scores [4]. Other commonly used input variables may include comorbidities, neurologic status, and biochemical parameters. For example, the preadmission comorbidities, level of consciousness, age, and neurological deficit (PLAN) score [5] includes comorbidities (preadmission dependence, cancer, congestive heart failure, and atrial fibrillation) and neurologic focal deficits (weakness of the leg or arm, aphasia, or neglect) as additional predictors. The Acute Stroke Registry and Analysis of Lausanne (ASTRAL) score [6] comprises age, stroke severity, stroke onset to admission time, the range of visual fields, acute glucose level, and the level of consciousness. However, the feasibility of these scores in daily clinical practice and their relevance to a specific clinical setting need to be well thought out prior to implementation [4]. Furthermore, using structured information alone, as well as the almost universal use of logistic regression models in the development of traditional prognostic scores [4,7], which require the assumption that linear and additive relationships are being fulfilled among predictors, significantly limits the applicability of these prognostic scores to an individual hospital or health system [8].

The ubiquitous use of electronic health records (EHRs) and the increase in computational power provide an opportunity to incorporate various types of structured data for the data-driven prediction of important clinical outcomes [9]. Machine learning algorithms have been used to develop prognostic models to predict various poststroke outcomes [10-16]. In previous studies that aimed to predict functional outcomes after an acute ischemic stroke (AIS), data-driven machine learning models generally performed equally as well as the PLAN and ASTRAL scores [10-12]. Matsumoto et al [10] developed and validated data-driven models via linear regression or decision tree ensembles and also validated traditional prognostic scores. Although no direct statistical comparisons of predictive performance were made between models, they concluded that

data-driven models may be alternative tools for predicting poststroke outcomes. Monteiro et al [11] found that machine learning models, including decision tree ensembles and support vector machines, achieved only a marginally higher predictive performance than that of traditional prognostic scores. Finally, Heo et al [12] found that machine learning models developed via random forest and logistic regression had a similar predictive performance to that of the ASTRAL score, while the deep neural network model outperformed this traditional prognostic score.

In addition to structured data, EHRs store a multitude of unstructured textual data, such as narrative clinical notes, radiology reports, and pathology reports. To our knowledge, this kind of information has not been explored in the development of stroke prognostic models [10-16]. However, natural language processing (NLP) has been used to extract valuable information stored in textual data within other medical applications. By harnessing the information from textual data, it is possible to improve the prognostication of patients with critical illness [8] and the detection of severe infection during emergency department triage [17]. Motivated by these ideas, we aimed to investigate whether machine learning models using unstructured clinical text can improve the prediction of functional outcomes at an early stage after AIS.

## Methods

### Study Settings

Data that support the study findings are available from the corresponding author on reasonable request. This retrospective study was conducted in a 1000-bed teaching hospital that had a catchment area with around 500,000 inhabitants. The stroke center of this hospital has been prospectively registering all patients who are hospitalized for a stroke and collecting data that conform to the design of the nationwide Taiwan Stroke Registry [18] since 2007. Data on patient demographics, personal and medical histories, stroke severity as assessed by using the National Institutes of Health Stroke Scale (NIHSS), the treatments that patients received, hospital courses, and final diagnoses were collected. Follow-up data, such as functional outcomes as assessed by the modified Rankin Scale (mRS), were collected only from patients who gave written informed consent for the follow-up evaluation.

### Ethics Approval

The study protocol was approved by the Ditmanson Medical Foundation Chia-Yi Christian Hospital Institutional Review Board (approval number: CYCH-IRB 2020090). Study data were maintained with confidentiality to ensure the privacy of all participants.

### Study Population

We identified all consecutive adult patients who were admitted to the study hospital for the first time for AIS from October 2007 to December 2019 by using the institutional stroke registry. Patients who experienced an in-hospital stroke or those who

were missing admission NIHSS scores from their clinical data were excluded. Those who did not provide consent for the follow-up or were lost to follow-up at 90 days were also excluded. For each patient, we retrieved the history of present illness (HPI) upon admission and the initial computed tomography (CT) report from the EHR database. Patients whose EHRs were unavailable were excluded.

To train and evaluate the machine learning models, we split the study population randomly into a training set that consisted of 75% (2885/3847) of the patients and a holdout test set that consisted of the remaining 25% (962/3847) of the patients, who were withheld from all models during the training process.

### Outcome Variable

The outcome of interest was a poor functional outcome as assessed by using the mRS score 90 days after a stroke. The mRS score was dichotomized into a good outcome (mRS score of 0-2) versus a poor outcome (mRS score of 3-6).

### Text Vectorization and Feature Selection

The model development and validation process is illustrated in [Figure 1](#). The free text extracted from the HPIs and CT reports was processed separately by using the following NLP techniques: (1) misspelled words were corrected by using the Jazzy spellchecker [19]; (2) abbreviations and acronyms were expanded to their full forms by looking up a list of common clinical abbreviations and acronyms, which is maintained by the stroke center of the study hospital ([Multimedia Appendix 1](#)); and (3) non-ASCII (American Standard Code for Information Interchange) characters and nonword symbols were removed.

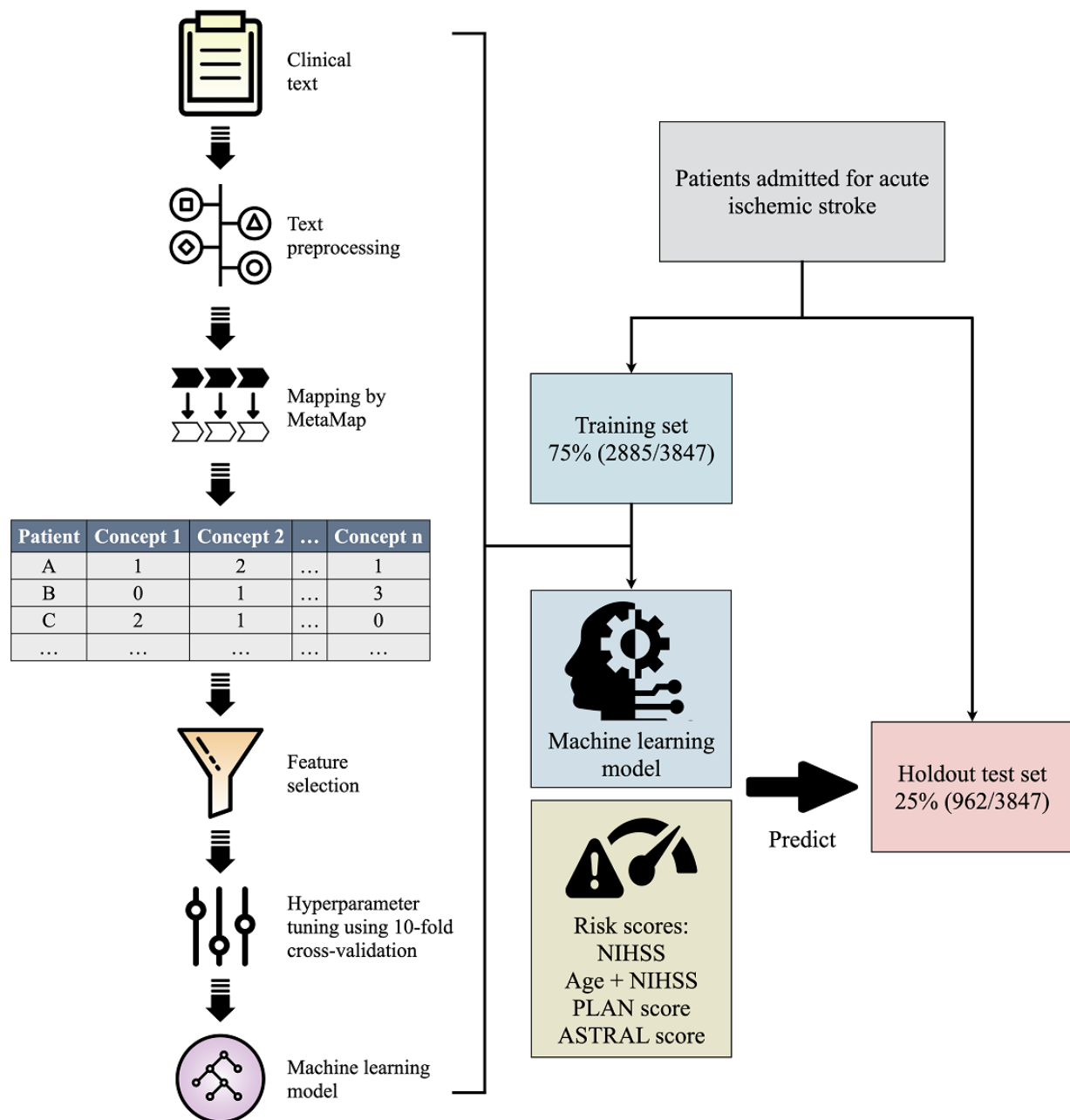
After text preprocessing, we used MetaMap to identify medical concepts from clinical text. MetaMap is an NLP tool that was

developed by the National Library of Medicine [20]. Through the process of tokenization, sentence boundary determination, part-of-speech tagging, and parsing, input text was decomposed and transformed to variants of words or phrases, which were mapped to medical concepts in the Unified Medical Language System Metathesaurus. MetaMap was configured with the option of using the NegEx algorithm to identify negated concepts. We appended the suffix *\_Neg* to concepts that were identified as negated. Next, the clinical text was vectorized for the text classification task by using the bag-of-words approach [21] or, more specifically, the so-called *bag-of-concepts approach* [22]. We built a document-term matrix in which each column represented each unique feature (concept) from the text corpus, the rows represented each document (the HPI or CT report for each patient), and the cells represented the counts of each concept within each document.

To reduce the number of redundant and less informative features and to improve training efficiency [21], we performed feature selection by filtering out concepts that appeared in less than 5% (145/2885) of all documents in the training set and then used 1 of the following 2 feature selection methods. The first method involved performing a penalized logistic regression with 10-fold cross-validation to identify the most predictive concepts [8,23]. The second involved using an extra tree classifier to determine important concepts based on the Gini index [24]. A large number of predictor variables (concepts) were still retained in the feature vector after these steps. To develop more parsimonious models, we built another document-term matrix by selecting the top 20 concepts that appeared in the documents of patients with poor or good functional outcomes based on chi-square statistics [25]. The same feature selection procedures were applied to the parsimonious models.



**Figure 1.** Model development and validation. ASTRAL: Acute Stroke Registry and Analysis of Lausanne; NIHSS: National Institutes of Health Stroke Scale; PLAN: preadmission comorbidities, level of consciousness, age, and neurological deficit.

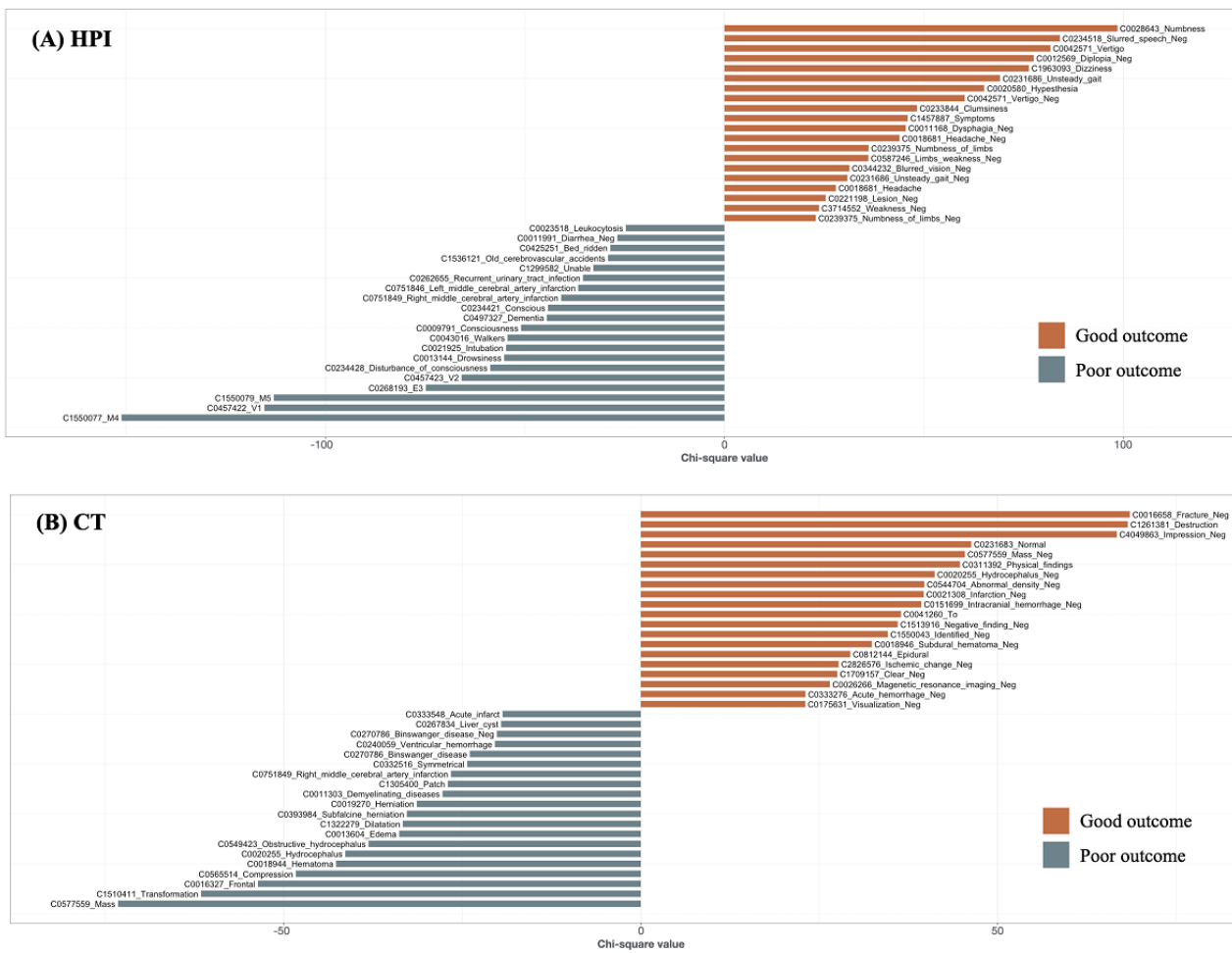


### Development of Machine Learning Models

Extreme gradient boosting (XGBoost) is an extension of gradient boosting algorithms [26]. It is an ensemble of classification and regression trees that can capture nonlinear interactions among input variables. The XGBoost algorithm trains a series of trees in which each subsequent tree attempts to correct the errors of the prior trees. XGBoost has gained popularity for predictive modeling in the medical field because of its high performance and scalability [24,27,28]. The XGBoost algorithm was implemented in Python 3.7 with xgboost Python package version 0.90.

We built 6 text-based models for predicting poor functional outcomes by using the XGBoost algorithm. Full model 1 was trained by using the features derived from the HPIs. Full model 2 was trained by using the features derived from both the HPIs and CT reports. In addition to the features used in full model 2, full model 3 included patient age as an input variable. Simple model 1 was trained by using only the selected concepts from the HPIs (Figure 2), and simple model 2 was trained by using the selected concepts from both the HPIs and CT reports (Figure 2). Similarly, simple model 3 included patient age.

**Figure 2.** Keyness plots showing the top 20 concepts that frequently appear in the (A) HPIs and (B) CT reports of patients with good or poor functional outcomes. The prefix before the concept is the concept unique identifier. A negated concept is suffixed with “\_Neg.” CT: computed tomography; HPI: history of present illness.



Hyperparameter optimization for each model was performed by repeatedly performing 10-fold cross-validation 10 times on the training set. We followed the steps proposed in a previous study [24] and conducted a grid search to find optimal hyperparameters. Model error was minimized in terms of the area under the receiver operating characteristic curve (AUC). Once the optimal hyperparameters were determined, the final models were fitted with the full training set.

With the introduction of machine learning techniques into health care settings, machine learning-based prediction models are being used to assist health care providers in decision-making for diagnosis, risk stratification, and clinical care. For decisions of such importance, clinicians prefer to know the reasons behind predictions rather than use a black-box model for prediction. The interpretability of model predictions is therefore considered a high priority for the implementation and use of prediction models [29]. To this end, after building the text-based models, we used Shapley additive explanations (SHAPs) [30], which are based on classic Shapley values from game theory, to explain the output of the XGBoost classifiers.

### Traditional Prognostic Models

A total of 4 traditional prognostic models based on the clinical data that were available at the time of admission were chosen

for experimentation. The model using NIHSS scores served as the first baseline model. The second baseline model consisted of age and NIHSS scores [3]. The third baseline model consisted of the PLAN scores [5]. The fourth baseline model consisted of the ASTRAL scores [6].

### Statistical Analysis

Categorical variables were expressed as counts and percentages, while continuous variables were expressed as means with SDs or medians with IQRs. Differences between groups were tested by using chi-square tests for categorical variables and 2-tailed *t* tests or Mann-Whitney *U* tests for continuous variables, as appropriate.

Model performance was evaluated on the test set. For each patient in the test set, the probability of a poor functional outcome was generated by using the six text-based machine learning models. To assess the predictive performance of each of the baseline models and text-based models, a logistic regression was used to predict a poor functional outcome. Furthermore, to assess the added usefulness of information from the clinical text, the output (the probability of a poor functional outcome) of simple model 2, which was based on unstructured clinical text from the HPIs and CT reports, was treated as an additional continuous variable and added to the baseline models.

Discriminatory ability was evaluated by calculating AUCs. The differences in AUCs among the models were compared by using the DeLong method [31]. In addition, improvements in predictive performance resulting from the addition of information from clinical text to each baseline model was evaluated by calculating the continuous net reclassification improvement and integrated discrimination improvement indices, as described by Pencina et al [32,33].

All statistical analyses were performed by using Stata 15.1 (StataCorp LLC) and R version 3.6.2 (R Foundation for Statistical Computing). Further, 2-tailed *P* values were considered statistically significant at  $<.05$ .

## Results

A total of 6176 patients were admitted for AIS. After excluding those with an in-hospital stroke ( $n=186$ ), those who were missing clinical data ( $n=216$ ), those who did not consent to the follow-up or were lost to follow-up ( $n=1048$ ), and those with unavailable EHRs ( $n=295$ ), the remaining 3847 patients comprised the study population. Of these, 1674 (43.5%) had a poor functional outcome after 90 poststroke days. Patients with a poor functional outcome were older, were more likely to be female, had more comorbidities (excluding hyperlipidemia), and were more likely to be dependent before the stroke. Stroke severity, PLAN scores, and ASTRAL scores were significantly higher among those with a poor functional outcome (Table 1).

**Table 1.** Baseline characteristics of the study population.

Characteristics	All (N=3847)	Functional outcome		<i>P</i> value
		Good (n=2173)	Poor (n=1674)	
Age (years), mean (SD)	69.5 (12.3)	66.1 (11.9)	74.0 (11.4)	<.001
Female, n (%)	1583 (41.1)	771 (35.5)	812 (48.5)	<.001
Hypertension, n (%)	3098 (80.5)	1694 (78)	1404 (83.9)	<.001
Diabetes mellitus, n (%)	1602 (41.6)	846 (38.9)	756 (45.2)	<.001
Hyperlipidemia, n (%)	2195 (57.1)	1323 (60.9)	872 (52.1)	<.001
Atrial fibrillation, n (%)	684 (17.8)	246 (11.3)	438 (26.2)	<.001
Congestive heart failure, n (%)	196 (5.1)	68 (3.1)	128 (7.6)	<.001
Cancer, n (%)	249 (6.5)	106 (4.9)	143 (8.5)	<.001
Preadmission dependence (mRS <sup>a</sup> score of >2), n (%)	419 (10.9)	29 (1.3)	390 (23.3)	<.001
Onset-to-admission delay (>3 hours), n (%)	2763 (71.8)	1574 (72.4)	1189 (71)	.34
NIHSS <sup>b</sup> score, median (IQR)	5 (3-10)	4 (2-6)	10 (5-19)	<.001
Glucose (mg/dl), mean (SD)	163 (83)	161 (82)	166 (84)	.06
PLAN <sup>c</sup> score, median (IQR)	8 (6-12)	7 (6-8)	12 (9-17)	<.001
ASTRAL <sup>d</sup> score, median (IQR)	21 (18-27)	19 (16-22)	27 (22-39)	<.001

<sup>a</sup>mRS: modified Rankin Scale.

<sup>b</sup>NIHSS: National Institutes of Health Stroke Scale.

<sup>c</sup>PLAN: preadmission comorbidities, level of consciousness, age, and neurological deficit.

<sup>d</sup>ASTRAL: Acute Stroke Registry and Analysis of Lausanne.

The training and test sets consisted of 2885 and 962 patients, respectively. The training set was used to build the document-term matrix and to train the machine learning models. Table S1 in Multimedia Appendix 2 lists the number of unique features and final selected features for each model. The AUCs of full models that used an extra tree classifier for feature selection were higher than the AUCs of those that used penalized logistic regression for feature selection, although the differences did not reach statistical significance. By contrast, penalized logistic regression resulted in higher AUCs than those resulting from extra tree classifiers for simple models, and a significant difference ( $P=.02$ ) was observed for simple model 3. Therefore, machine learning models that used penalized logistic regression for feature selection were used in the following analyses.

The top 20 features for both good and poor functional outcomes that were used in the simple models are shown in Figure 2. Figure 3 shows the top 20 most important text features from simple model 2; the features are ordered by the average absolute SHAP value, which indicates the magnitude of the impact on model output. Figure 3 also presents bee swarm plots showing the magnitude and direction of the effect of each feature according to the SHAP value, demonstrating how simple model 2 uses input features to make predictions. For example, when the concepts of *symmetrical*, *Binswanger disease*, or *dilatation* appear in a CT report, the model tends to predict a poor outcome, whereas the model tends to predict a good outcome when an HPI contains the concepts of *numbness* or the negated form of *slurred speech*. Figures S1-S6 in Multimedia Appendix 2 show the bee swarm plots for all text-based models.

**Figure 3.** (A) A bar chart showing the top 20 most important features of simple model 2 according to the average absolute SHAP values, which indicate the average impact on model output. (B) A bee swarm plot for the top 20 features in which each dot represents an individual patient. A dot's position on the x-axis indicates the impact that a feature has on the model's prediction for that patient. The color of the dot specifies the relative value of the corresponding feature (concept). A higher feature value means that the concept appears more times in the clinical text. The prefix before the concept is the concept unique identifier. A negated concept is suffixed with “\_Neg”. CT: computed tomography; HPI: history of present illness; SHAP: Shapley additive explanations.

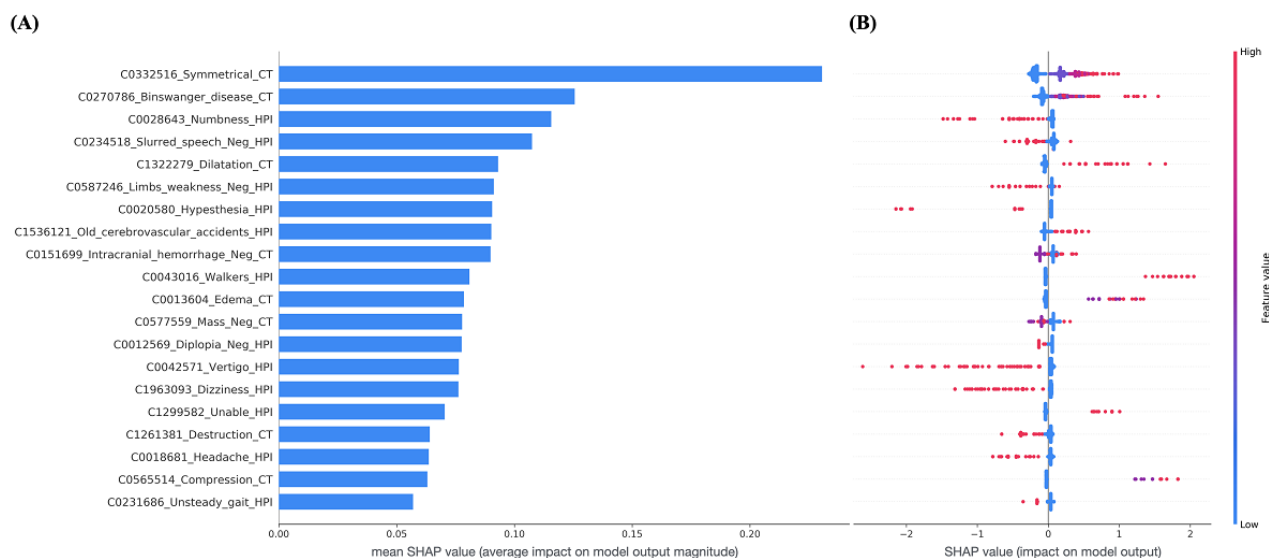
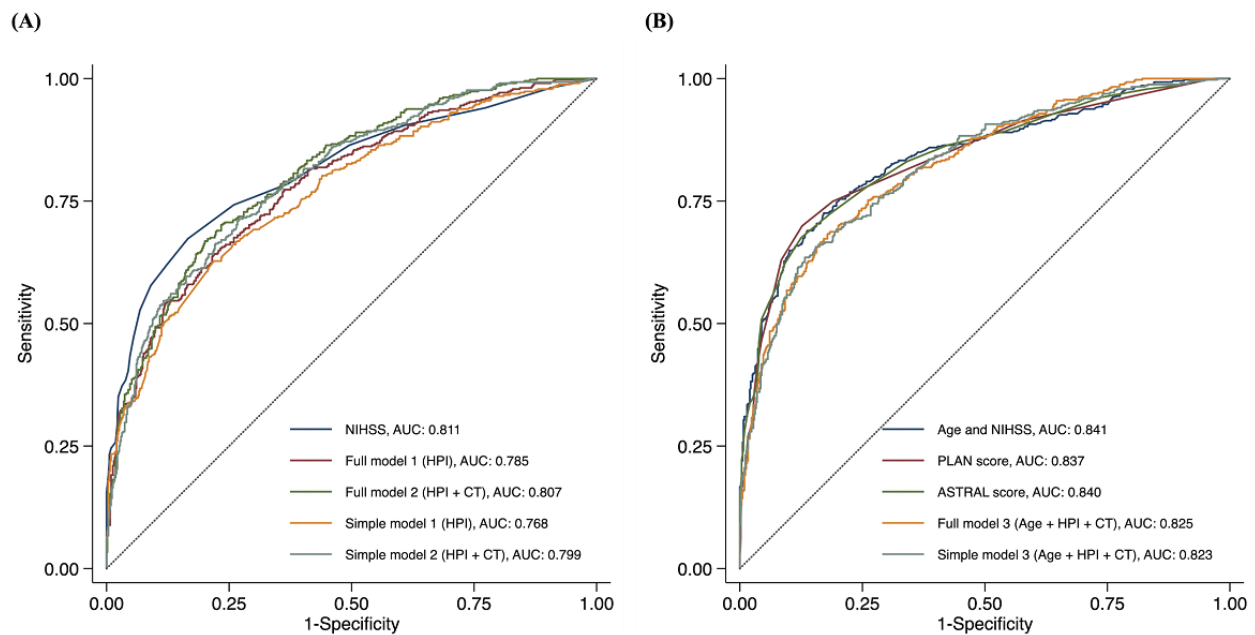


Figure 4 illustrates the receiver operating characteristic curves for the six text-based models and the four baseline models trained on the test set. The models are grouped according to whether age is included in the model. Tables S2-S4 in Multimedia Appendix 2 list these models' AUCs (with 95% CIs) and the *P* values for the pairwise comparison of model performance. Models that included age generally had higher AUC values (range 0.823-0.841) than those of the models that did not include age (range 0.768-0.811). Among the models that did not include age, the AUCs of full model 1 (0.785; 95% CI 0.756-0.814), full model 2 (0.807; 95% CI 0.779-0.834), and simple model 2 (0.799; 95% CI 0.771-0.827) were not significantly different from that of the model that included NIHSS scores (0.811; 95% CI 0.783-0.839; *P* = .11, .78, and .47, respectively). Among the models that included age, the AUCs of full model 3 (0.825; 95% CI 0.799-0.851) and simple model 3 (0.823; 95% CI 0.797-0.850) were also not significantly

different from those of the model that included age and NIHSS scores (0.841; 95% CI 0.815-0.867; *P* = .22 and .17, respectively), the model that included the PLAN scores (0.837; 95% CI 0.811-0.863; *P* = .37 and .30, respectively), and the model that included the ASTRAL scores (0.840; 95% CI 0.814-0.866; *P* = .27 and .22, respectively). Table 2 lists the predictive performance of models with and without added information from the clinical text. According to the AUCs (model including age, NIHSS scores, and text: *P* = .002; model include PLAN scores and text: *P* < .001; model including ASTRAL scores and text: *P* = .004), net reclassification improvement indices (all models including text: *P* < .001), and integrated discrimination improvement indices (all models including text: *P* < .001), a statistically significant improvement in predictive performance was achieved when adding information from the clinical text into the baseline models.

**Figure 4.** Receiver operating characteristic curves for predicting a poor functional outcome for (A) models without age and (B) models with age. ASTRAL: Acute Stroke Registry and Analysis of Lausanne; AUC: area under the receiver operating characteristic curve; CT: computed tomography; HPI: history of present illness; NIHSS: National Institutes of Health Stroke Scale; PLAN: preadmission comorbidities, level of consciousness, age, and neurological deficit.



**Table 2.** Comparison of the performance of baseline models with or without added information from clinical text.

Model	AUC <sup>a</sup> (95% CI)	P value	NRI <sup>b</sup> , % (95% CI)	P value	IDI <sup>c</sup> , % (95% CI)	P value
Age and NIHSS <sup>d</sup> score	0.841 (0.815-0.867)	N/A <sup>e</sup>	N/A	N/A	N/A	N/A
Age and NIHSS score plus text	0.861 (0.837-0.885)	.002	0.427 (0.302-0.551)	<.001	0.042 (0.029-0.054)	<.001
PLAN <sup>f</sup> score	0.837 (0.811-0.863)	N/A	N/A	N/A	N/A	N/A
PLAN score plus text	0.856 (0.835-0.882)	<.001	0.543 (0.420-0.665)	<.001	0.038 (0.026-0.051)	<.001
ASTRAL <sup>g</sup> score	0.840 (0.814-0.866)	N/A	N/A	N/A	N/A	N/A
ASTRAL score plus text	0.860 (0.837-0.884)	.004	0.443 (0.318-0.567)	<.001	0.044 (0.031-0.057)	<.001

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

<sup>b</sup>NRI: net reclassification improvement.

<sup>c</sup>IDI: integrated discrimination improvement.

<sup>d</sup>NIHSS: National Institutes of Health Stroke Scale.

<sup>e</sup>N/A: not applicable.

<sup>f</sup>PLAN: preadmission comorbidities, level of consciousness, age, and neurological deficit.

<sup>g</sup>ASTRAL indicates Acute Stroke Registry and Analysis of Lausanne.

## Discussion

### Principal Findings

This study demonstrates that machine learning models based on clinical text may provide an alternative way of prognosticating patients after AIS. Most of the models (3/4, 75%) based on textual data alone performed equally as well as the models based on NIHSS scores, whereas models based on text and patient age had a comparable predictive performance to those of the model based on age and NIHSS scores, the model based on the PLAN scores, and the model based on the ASTRAL scores. In addition, the information extracted from clinical text can be used to improve the predictive performance of existing

prognostic scores in terms of the prediction of the 90-day functional outcome.

Previous studies have found that machine learning algorithms had comparable discrimination to or even higher discrimination than that of conventional logistic regression models [10-12]. A possible explanation may be that machine learning algorithms can capture potential nonlinear relationships and handle complex interactions between the input variables and the outcome variable [10,34,35]. On the other hand, the performance of prognostic scores is generally limited by different demographic and risk factor distributions across diverse populations and health care settings [36,37]. By contrast, data-driven models can make predictions without prior knowledge of the real system

[38]. The use of machine learning methods may enable each individual site to develop its own prediction models for providing patients with individualized medical decisions and treatments. However, their transferability to different health systems is not guaranteed.

Despite the emergence of machine learning technology as a new tool for prognosticating stroke outcomes, textual data have rarely been analyzed or used in previous machine learning prediction models in the field of stroke medicine [39-44]. By using NLP techniques, information extracted from unstructured text, such as clinical notes or radiology reports, has been used to build machine learning models to identify AIS [39-41] or automate AIS subtype classification [43,44]. One of the advantages of using textual data is that narrative notes are generated during routine health care processes, thus avoiding the extra effort required for data collection and coding. Although structured entry and reporting tools are now available for clinical documentation, health care providers generally prefer to write narrative notes because structured documentation systems can be too awkward to use without impeding clinical workflows and can even result in errors [45,46]. Furthermore, the excessive use of structured data entry in clinical documentation tends to result in the loss of the subtleties in information by standardizing away the heterogeneity across patients [46].

Although only the basic bag-of-words model was used for text representation, this study shows an application of text classification in the development of clinical prediction models. However, a major challenge of this approach is the high dimensionality of the feature space. The large number of features generated by the bag-of-words model may cause problems, such as increased computational complexity, degraded classification performance, and overfitting [21,47]. Feature selection is thus a necessary step for text classification. However, the choice of feature selection methods usually depends on the characteristics of the data and requires trade-offs among multiple criteria, particularly in small samples with high dimensionality [47]. According to our experiments, the two feature selection methods indeed performed slightly differently in different situations.

Another merit of using the bag-of-words approach for text vectorization is the high level of interpretability that can be achieved; this approach allows domain experts to examine each predictor (concept) within its specific context. The patterns that a machine learning model discovers and the explanations for what is observed can be more important than the model's predictive performance, particularly in medical applications. In

this regard, we applied Shapley values to measure the impact of each predictor. Taking the concept *symmetrical* as an example, the reason why this concept tends to be associated with a poor functional outcome (Figure 3) may not be obvious at first glance. The reason became clear when the original text in the CT reports was reviewed. Radiologists generally described subcortical arteriosclerotic encephalopathy as “symmetrical hypodensities in bilateral periventricular regions” and mentioned hydrocephalus as a “symmetrical enlargement of the lateral ventricles.” Both conditions cause a range of impairments in brain function. Consequently, the concept *symmetrical* is commonly found in the CT reports of patients with a poor functional outcome.

### Limitations

This study had some limitations to be addressed. First, although data-driven prediction approaches have their own merits, the relationships discovered from our data do not necessarily indicate causation; therefore, prediction accuracy should never be interpreted as causal validity [48]. Second, this is a single-site study, which may limit the generalizability of study results. Third, although MetaMap was used to extract medical concepts, this study basically adopted the bag-of-words approach to represent clinical text. As such, it disregards the order of concepts and does not capture the contextual dependency between concepts. Furthermore, different kinds of speculative expressions, ranging from completely affirmative to completely nonaffirmative, were found in the clinical text. Even though negation detection was used, we did not perform factuality detection. Different types of text representations, such as contextual word embeddings, may be explored in future research. Fourth, the terms and phrases used in clinical documentation may differ across health systems and cultures. This renders the transferability of the machine learning models questionable and may entail that each individual health system has to build its own version of the prediction models and follow a similar process of model development.

### Conclusions

This study demonstrates that by using NLP and machine learning techniques, unstructured clinical text has the potential to improve the early prediction of functional outcomes after AIS. Despite these findings, this does not mean that the machine learning models developed in this study can be directly deployed at other stroke centers. We further suggest that each individual health system develops its own model by applying the proposed methods to its EHRs.

---

### Acknowledgments

The authors would like to thank Ms Li-Ying Sung for providing English language editing support. This research was funded by the Ditmanson Medical Foundation Chia-Yi Christian Hospital Research Program (grant R109-37-1). The funder of the research had no role in the design and conduct of the study, interpretation of the data, or decision to submit for publication.

---

### Conflicts of Interest

None declared.

---

## Multimedia Appendix 1

List of common clinical abbreviations and acronyms.

[\[XLSX File \(Microsoft Excel File\), 18 KB - medinform\\_v10i2e29806\\_app1.xlsx \]](#)

## Multimedia Appendix 2

Supplemental material.

[\[PDF File \(Adobe PDF File\), 3455 KB - medinform\\_v10i2e29806\\_app2.pdf \]](#)

## References

1. GBD 2016 Lifetime Risk of Stroke Collaborators, Feigin VL, Nguyen G, Cercy K, Johnson CO, Alam T, et al. Global, regional, and country-specific lifetime risks of stroke, 1990 and 2016. *N Engl J Med* 2018 Dec 20;379(25):2429-2437 [FREE Full text] [doi: [10.1056/NEJMoa1804492](https://doi.org/10.1056/NEJMoa1804492)] [Medline: [30575491](https://pubmed.ncbi.nlm.nih.gov/30575491/)]
2. Campbell BCV, Khatri P. Stroke. *Lancet* 2020 Jul 11;396(10244):129-142. [doi: [10.1016/S0140-6736\(20\)31179-X](https://doi.org/10.1016/S0140-6736(20)31179-X)] [Medline: [32653056](https://pubmed.ncbi.nlm.nih.gov/32653056/)]
3. Weimar C, König IR, Kraywinkel K, Ziegler A, Diener HC, German Stroke Study Collaboration. Age and National Institutes of Health Stroke Scale Score within 6 hours after onset are accurate predictors of outcome after cerebral ischemia: development and external validation of prognostic models. *Stroke* 2004 Jan;35(1):158-162. [doi: [10.1161/01.STR.0000106761.94985.8B](https://doi.org/10.1161/01.STR.0000106761.94985.8B)] [Medline: [14684776](https://pubmed.ncbi.nlm.nih.gov/14684776/)]
4. Drozdowska BA, Singh S, Quinn TJ. Thinking about the future: A review of prognostic scales used in acute stroke. *Front Neurol* 2019 Mar 21;10:274 [FREE Full text] [doi: [10.3389/fneur.2019.00274](https://doi.org/10.3389/fneur.2019.00274)] [Medline: [30949127](https://pubmed.ncbi.nlm.nih.gov/30949127/)]
5. O'Donnell MJ, Fang J, D'Uva C, Saposnik G, Gould L, McGrath E, Investigators of the Registry of the Canadian Stroke Network. The PLAN score: a bedside prediction rule for death and severe disability following acute ischemic stroke. *Arch Intern Med* 2012 Nov 12;172(20):1548-1556. [doi: [10.1001/2013.jamainternmed.30](https://doi.org/10.1001/2013.jamainternmed.30)] [Medline: [23147454](https://pubmed.ncbi.nlm.nih.gov/23147454/)]
6. Ntaios G, Faouzi M, Ferrari J, Lang W, Vemmos K, Michel P. An integer-based score to predict functional outcome in acute ischemic stroke: the ASTRAL score. *Neurology* 2012 Jun 12;78(24):1916-1922. [doi: [10.1212/WNL.0b013e318259e221](https://doi.org/10.1212/WNL.0b013e318259e221)] [Medline: [22649218](https://pubmed.ncbi.nlm.nih.gov/22649218/)]
7. Christodoulou E, Ma J, Collins GS, Steyerberg EW, Verbakel JY, Van Calster B. A systematic review shows no performance benefit of machine learning over logistic regression for clinical prediction models. *J Clin Epidemiol* 2019 Jun;110:12-22. [doi: [10.1016/j.jclinepi.2019.02.004](https://doi.org/10.1016/j.jclinepi.2019.02.004)] [Medline: [30763612](https://pubmed.ncbi.nlm.nih.gov/30763612/)]
8. Weissman GE, Hubbard RA, Ungar LH, Harhay MO, Greene CS, Himes BE, et al. Inclusion of unstructured clinical text improves early prediction of death or prolonged ICU stay. *Crit Care Med* 2018 Jul;46(7):1125-1132 [FREE Full text] [doi: [10.1097/CCM.0000000000003148](https://doi.org/10.1097/CCM.0000000000003148)] [Medline: [29629986](https://pubmed.ncbi.nlm.nih.gov/29629986/)]
9. Ding L, Liu C, Li Z, Wang Y. Incorporating artificial intelligence into stroke care and research. *Stroke* 2020 Dec;51(12):e351-e354 [FREE Full text] [doi: [10.1161/STROKEAHA.120.031295](https://doi.org/10.1161/STROKEAHA.120.031295)] [Medline: [33106108](https://pubmed.ncbi.nlm.nih.gov/33106108/)]
10. Matsumoto K, Nohara Y, Soejima H, Yonehara T, Nakashima N, Kamouchi M. Stroke prognostic scores and data-driven prediction of clinical outcomes after acute ischemic stroke. *Stroke* 2020 May;51(5):1477-1483. [doi: [10.1161/STROKEAHA.119.027300](https://doi.org/10.1161/STROKEAHA.119.027300)] [Medline: [32208843](https://pubmed.ncbi.nlm.nih.gov/32208843/)]
11. Monteiro M, Fonseca AC, Freitas AT, Melo TPE, 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 Bioinform* 2018;15(6):1953-1959. [doi: [10.1109/TCBB.2018.2811471](https://doi.org/10.1109/TCBB.2018.2811471)] [Medline: [29994736](https://pubmed.ncbi.nlm.nih.gov/29994736/)]
12. 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](https://doi.org/10.1161/STROKEAHA.118.024293)] [Medline: [30890116](https://pubmed.ncbi.nlm.nih.gov/30890116/)]
13. Xie Y, Jiang B, Gong E, Li Y, Zhu G, Michel P, et al. JOURNAL CLUB: Use of gradient boosting machine Learning to predict patient outcome in acute ischemic stroke on the basis of imaging, demographic, and clinical information. *AJR Am J Roentgenol* 2019 Jan;212(1):44-51. [doi: [10.2214/AJR.18.20260](https://doi.org/10.2214/AJR.18.20260)] [Medline: [30354266](https://pubmed.ncbi.nlm.nih.gov/30354266/)]
14. Li X, Pan X, Jiang C, Wu M, Liu Y, Wang F, et al. Predicting 6-month unfavorable outcome of acute ischemic stroke using machine learning. *Front Neurol* 2020 Nov 19;11:539509 [FREE Full text] [doi: [10.3389/fneur.2020.539509](https://doi.org/10.3389/fneur.2020.539509)] [Medline: [33329298](https://pubmed.ncbi.nlm.nih.gov/33329298/)]
15. Alaka SA, Menon BK, Brobbey A, Williamson T, Goyal M, Demchuk AM, et al. Functional outcome prediction in ischemic stroke: A comparison of machine learning algorithms and regression models. *Front Neurol* 2020 Aug 25;11:889 [FREE Full text] [doi: [10.3389/fneur.2020.00889](https://doi.org/10.3389/fneur.2020.00889)] [Medline: [32982920](https://pubmed.ncbi.nlm.nih.gov/32982920/)]
16. Lin CH, Hsu KC, Johnson KR, Fann YC, Tsai CH, Sun Y, Taiwan Stroke Registry Investigators. Evaluation of machine learning methods to stroke outcome prediction using a nationwide disease registry. *Comput Methods Programs Biomed* 2020 Jul;190:105381 [FREE Full text] [doi: [10.1016/j.cmpb.2020.105381](https://doi.org/10.1016/j.cmpb.2020.105381)] [Medline: [32044620](https://pubmed.ncbi.nlm.nih.gov/32044620/)]
17. Horng S, Sontag DA, Halpern Y, Jernite Y, Shapiro NI, Nathanson LA. Creating an automated trigger for sepsis clinical decision support at emergency department triage using machine learning. *PLoS One* 2017 Apr 06;12(4):e0174708. [doi: [10.1371/journal.pone.0174708](https://doi.org/10.1371/journal.pone.0174708)] [Medline: [28384212](https://pubmed.ncbi.nlm.nih.gov/28384212/)]

18. Hsieh FI, Lien LM, Chen ST, Bai CH, Sun MC, Tseng HP, Taiwan Stroke Registry Investigators. Get with the guidelines-Stroke performance indicators: surveillance of stroke care in the Taiwan Stroke Registry: Get with the guidelines-stroke in Taiwan. *Circulation* 2010 Sep 14;122(11):1116-1123. [doi: [10.1161/CIRCULATIONAHA.110.936526](https://doi.org/10.1161/CIRCULATIONAHA.110.936526)] [Medline: [20805428](https://pubmed.ncbi.nlm.nih.gov/20805428/)]
19. Idzelis M. The Java Open Source Spell Checker. SourceForge. URL: <http://jazzy.sourceforge.net/> [accessed 2021-07-03]
20. Aronson AR, Lang FM. An overview of MetaMap: historical perspective and recent advances. *J Am Med Inform Assoc* 2010;17(3):229-236 [FREE Full text] [doi: [10.1136/jamia.2009.002733](https://doi.org/10.1136/jamia.2009.002733)] [Medline: [20442139](https://pubmed.ncbi.nlm.nih.gov/20442139/)]
21. Deng X, Li Y, Weng J, Zhang J. Feature selection for text classification: A review. *Multimed Tools Appl* 2018 May 8;78(3):3797-3816. [doi: [10.1007/s11042-018-6083-5](https://doi.org/10.1007/s11042-018-6083-5)]
22. Mujtaba G, Shuib L, Idris N, Hoo WL, Raj RG, Khowaja K, et al. Clinical text classification research trends: Systematic literature review and open issues. *Expert Syst Appl* 2019 Feb;116:494-520. [doi: [10.1016/j.eswa.2018.09.034](https://doi.org/10.1016/j.eswa.2018.09.034)]
23. Ma S, Huang J. Penalized feature selection and classification in bioinformatics. *Brief Bioinform* 2008 Sep;9(5):392-403 [FREE Full text] [doi: [10.1093/bib/bbn027](https://doi.org/10.1093/bib/bbn027)] [Medline: [18562478](https://pubmed.ncbi.nlm.nih.gov/18562478/)]
24. Ogunleye A, Wang QG. XGBoost model for chronic kidney disease diagnosis. *IEEE/ACM Trans Comput Biol Bioinform* 2020;17(6):2131-2140. [doi: [10.1109/TCBB.2019.2911071](https://doi.org/10.1109/TCBB.2019.2911071)] [Medline: [30998478](https://pubmed.ncbi.nlm.nih.gov/30998478/)]
25. Culpeper J. Keyness: Words, parts-of-speech and semantic categories in the character-talk of Shakespeare's Romeo and Juliet. *International Journal of Corpus Linguistics* 2009 Jan;14(1):29-59. [doi: [10.1075/ijcl.14.1.03cul](https://doi.org/10.1075/ijcl.14.1.03cul)]
26. Chen T, Guestrin C. XGBoost: A scalable tree boosting system. 2016 Aug Presented at: KDD '16: The 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining; August 13-17, 2016; San Francisco, California, USA p. 785-794. [doi: [10.1145/2939672.2939785](https://doi.org/10.1145/2939672.2939785)]
27. Xu Y, Yang X, Huang H, Peng C, Ge Y, Wu H, et al. Extreme gradient boosting model has a better performance in predicting the risk of 90-day readmissions in patients with ischaemic stroke. *J Stroke Cerebrovasc Dis* 2019 Dec;28(12):104441. [doi: [10.1016/j.jstrokecerebrovasdis.2019.104441](https://doi.org/10.1016/j.jstrokecerebrovasdis.2019.104441)] [Medline: [31627995](https://pubmed.ncbi.nlm.nih.gov/31627995/)]
28. Shimoda A, Ichikawa D, Oyama H. Using machine-learning approaches to predict non-participation in a nationwide general health check-up scheme. *Comput Methods Programs Biomed* 2018 Sep;163:39-46. [doi: [10.1016/j.cmpb.2018.05.032](https://doi.org/10.1016/j.cmpb.2018.05.032)] [Medline: [30119856](https://pubmed.ncbi.nlm.nih.gov/30119856/)]
29. Ahmad MA, Eckert C, Teredesai A. Interpretable machine learning in healthcare. 2018 Aug Presented at: BCB '18: 9th ACM International Conference on Bioinformatics, Computational Biology and Health Informatics; August 29 to September 1, 2018; Washington, DC, USA p. 559-560. [doi: [10.1145/3233547.3233667](https://doi.org/10.1145/3233547.3233667)]
30. Lundberg SM, Erion G, Chen H, DeGrave A, Prutkin JM, Nair B, et al. From local explanations to global understanding with explainable AI for trees. *Nat Mach Intell* 2020 Jan;2(1):56-67 [FREE Full text] [doi: [10.1038/s42256-019-0138-9](https://doi.org/10.1038/s42256-019-0138-9)] [Medline: [32607472](https://pubmed.ncbi.nlm.nih.gov/32607472/)]
31. 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](https://doi.org/10.2307/2531595)]
32. Pencina MJ, D'Agostino Sr RB, D'Agostino Jr RB, Vasan RS. Evaluating the added predictive ability of a new marker: from area under the ROC curve to reclassification and beyond. *Stat Med* 2008 Jan 30;27(2):157-72; discussion 207-212. [doi: [10.1002/sim.2929](https://doi.org/10.1002/sim.2929)] [Medline: [17569110](https://pubmed.ncbi.nlm.nih.gov/17569110/)]
33. Pencina MJ, D'Agostino Sr RB, Steyerberg EW. Extensions of net reclassification improvement calculations to measure usefulness of new biomarkers. *Stat Med* 2011 Jan 15;30(1):11-21 [FREE Full text] [doi: [10.1002/sim.4085](https://doi.org/10.1002/sim.4085)] [Medline: [21204120](https://pubmed.ncbi.nlm.nih.gov/21204120/)]
34. Orfanoudaki A, Chesley E, Cadisch C, Stein B, Nouh A, Alberts MJ, et al. Machine learning provides evidence that stroke risk is not linear: The non-linear Framingham stroke risk score. *PLoS One* 2020 May 21;15(5):e0232414. [doi: [10.1371/journal.pone.0232414](https://doi.org/10.1371/journal.pone.0232414)] [Medline: [32437368](https://pubmed.ncbi.nlm.nih.gov/32437368/)]
35. van Os HJA, Ramos LA, Hilbert A, van Leeuwen M, van Walderveen MAA, Kruijff ND, MR CLEAN Registry Investigators. Predicting outcome of endovascular treatment for acute ischemic stroke: Potential value of machine learning algorithms. *Front Neurol* 2018 Sep 25;9:784 [FREE Full text] [doi: [10.3389/fneur.2018.00784](https://doi.org/10.3389/fneur.2018.00784)] [Medline: [30319525](https://pubmed.ncbi.nlm.nih.gov/30319525/)]
36. Glümer C, Vistisen D, Borch-Johnsen K, Colagiuri S, DETECT-2 Collaboration. Risk scores for type 2 diabetes can be applied in some populations but not all. *Diabetes Care* 2006 Feb;29(2):410-414. [doi: [10.2337/diacare.29.02.06.dc05-0945](https://doi.org/10.2337/diacare.29.02.06.dc05-0945)] [Medline: [16443896](https://pubmed.ncbi.nlm.nih.gov/16443896/)]
37. Quinn GR, Severdija ON, Chang Y, Singer DE. Wide variation in reported rates of stroke across cohorts of patients with atrial fibrillation. *Circulation* 2017 Jan 17;135(3):208-219. [doi: [10.1161/CIRCULATIONAHA.116.024057](https://doi.org/10.1161/CIRCULATIONAHA.116.024057)] [Medline: [27799272](https://pubmed.ncbi.nlm.nih.gov/27799272/)]
38. Alaa AM, Bolton T, Di Angelantonio E, Rudd JHF, van der Schaar M. Cardiovascular disease risk prediction using automated machine learning: A prospective study of 423,604 UK Biobank participants. *PLoS One* 2019 May 15;14(5):e0213653. [doi: [10.1371/journal.pone.0213653](https://doi.org/10.1371/journal.pone.0213653)] [Medline: [31091238](https://pubmed.ncbi.nlm.nih.gov/31091238/)]
39. Sedghi E, Weber JH, Thomo A, Bibok M, Penn AMW. Mining clinical text for stroke prediction. *Netw Model Anal Health Inform Bioinform* 2015 Jul 14;4(16):688. [doi: [10.1007/s13721-015-0090-5](https://doi.org/10.1007/s13721-015-0090-5)]



40. Kim C, Zhu V, Obeid J, Lenert L. Natural language processing and machine learning algorithm to identify brain MRI reports with acute ischemic stroke. *PLoS One* 2019 Feb 28;14(2):e0212778. [doi: [10.1371/journal.pone.0212778](https://doi.org/10.1371/journal.pone.0212778)] [Medline: [30818342](https://pubmed.ncbi.nlm.nih.gov/30818342/)]
41. Ong CJ, Orfanoudaki A, Zhang R, Caprasse FPM, Hutch M, Ma L, et al. Machine learning and natural language processing methods to identify ischemic stroke, acuity and location from radiology reports. *PLoS One* 2020 Jun 19;15(6):e0234908. [doi: [10.1371/journal.pone.0234908](https://doi.org/10.1371/journal.pone.0234908)] [Medline: [32559211](https://pubmed.ncbi.nlm.nih.gov/32559211/)]
42. Govindarajan P, Soundarapandian RK, Gandomi AH, Patan R, Jayaraman P, Manikandan R. Classification of stroke disease using machine learning algorithms. *Neural Comput Appl* 2019 Jan 25;32:817-828. [doi: [10.1007/s00521-019-04041-y](https://doi.org/10.1007/s00521-019-04041-y)]
43. Garg R, Oh E, Naidech A, Kording K, Prabhakaran S. Automating ischemic stroke subtype classification using machine learning and natural language processing. *J Stroke Cerebrovasc Dis* 2019 Jul;28(7):2045-2051. [doi: [10.1016/j.jstrokecerebrovasdis.2019.02.004](https://doi.org/10.1016/j.jstrokecerebrovasdis.2019.02.004)] [Medline: [31103549](https://pubmed.ncbi.nlm.nih.gov/31103549/)]
44. Sung SF, Lin CY, Hu YH. EMR-based phenotyping of ischemic stroke using supervised machine learning and text mining techniques. *IEEE J Biomed Health Inform* 2020 Oct;24(10):2922-2931. [doi: [10.1109/JBHI.2020.2976931](https://doi.org/10.1109/JBHI.2020.2976931)] [Medline: [32142458](https://pubmed.ncbi.nlm.nih.gov/32142458/)]
45. Rosenbloom ST, Denny JC, Xu H, Lorenzi N, Stead WW, Johnson KB. Data from clinical notes: a perspective on the tension between structure and flexible documentation. *J Am Med Inform Assoc* 2011;18(2):181-186 [FREE Full text] [doi: [10.1136/jamia.2010.007237](https://doi.org/10.1136/jamia.2010.007237)] [Medline: [21233086](https://pubmed.ncbi.nlm.nih.gov/21233086/)]
46. Kuhn T, Basch P, Barr M, Yackel T, Medical Informatics Committee of the American College of Physicians. Clinical documentation in the 21st century: executive summary of a policy position paper from the American College of Physicians. *Ann Intern Med* 2015 Feb 17;162(4):301-303 [FREE Full text] [doi: [10.7326/M14-2128](https://doi.org/10.7326/M14-2128)] [Medline: [25581028](https://pubmed.ncbi.nlm.nih.gov/25581028/)]
47. Kou G, Yang P, Peng Y, Xiao F, Chen Y, Alsaadi FE. Evaluation of feature selection methods for text classification with small datasets using multiple criteria decision-making methods. *Appl Soft Comput* 2020 Jan;86:105836 [FREE Full text] [doi: [10.1016/j.asoc.2019.105836](https://doi.org/10.1016/j.asoc.2019.105836)]
48. Li J, Liu L, Le TD, Liu J. Accurate data-driven prediction does not mean high reproducibility. *Nat Mach Intell* 2020 Jan 17;2:13-15. [doi: [10.1038/s42256-019-0140-2](https://doi.org/10.1038/s42256-019-0140-2)]

## Abbreviations

**AIS:** acute ischemic stroke

**ASCII:** American Standard Code for Information Interchange

**ASTRAL:** Acute Stroke Registry and Analysis of Lausanne

**AUC:** area under the receiver operating characteristic curve

**CT:** computed tomography

**EHR:** electronic health record

**HPI:** history of present illness

**mRS:** modified Rankin Scale

**NIHSS:** National Institutes of Health Stroke Scale

**NLP:** natural language processing

**PLAN:** preadmission comorbidities, level of consciousness, age, and neurological deficit

**SHAP:** Shapley additive explanation

**XGBoost:** extreme gradient boosting

*Edited by C Lovis; submitted 28.04.21; peer-reviewed by C Huang, M Burns; comments to author 28.06.21; revised version received 17.07.21; accepted 02.01.22; published 17.02.22.*

*Please cite as:*

Sung SF, Hsieh CY, Hu YH

Early Prediction of Functional Outcomes After Acute Ischemic Stroke Using Unstructured Clinical Text: Retrospective Cohort Study  
*JMIR Med Inform* 2022;10(2):e29806

URL: <https://medinform.jmir.org/2022/2/e29806>

doi: [10.2196/29806](https://doi.org/10.2196/29806)

PMID: [35175201](https://pubmed.ncbi.nlm.nih.gov/35175201/)

©Sheng-Feng Sung, Cheng-Yang Hsieh, Ya-Han Hu. Originally published in *JMIR Medical Informatics* (<https://medinform.jmir.org>), 17.02.2022. This is an open-access article distributed under the terms of the Creative Commons Attribution License (<https://creativecommons.org/licenses/by/4.0/>), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in *JMIR Medical Informatics*, is properly cited. The complete

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

Original Paper

# The Application and Comparison of Machine Learning Models for the Prediction of Breast Cancer Prognosis: Retrospective Cohort Study

Jialong Xiao<sup>1,2,3</sup>, BS, MS; Miao Mo<sup>2,3</sup>, BS, MS; Zezhou Wang<sup>2,3</sup>, BS, MS; Changming Zhou<sup>2,3</sup>, BS, MS, PhD; Jie Shen<sup>2,3</sup>, BS, MS, PhD; Jing Yuan<sup>2,3</sup>, BS; Yulian He<sup>1,2</sup>, BS, MS; Ying Zheng<sup>2,3,4</sup>, BS, MS

<sup>1</sup>Department of Epidemiology, School of Public Health, Fudan University, Shanghai, China

<sup>2</sup>Department of Cancer Prevention, Fudan University Shanghai Cancer Center, Shanghai, China

<sup>3</sup>Department of Oncology, Shanghai Medical College, Fudan University, Shanghai, China

<sup>4</sup>Shanghai Engineering Research Center of Artificial Intelligence Technology for Tumor Diseases, Shanghai, China

**Corresponding Author:**

Ying Zheng, BS, MS

Department of Cancer Prevention

Fudan University Shanghai Cancer Center

270 Dong 'an Road, Xuhui District

Shanghai, 200000

China

Phone: 86 21 64175590

Fax: 86 21 64175590

Email: [zhengying@fudan.edu.cn](mailto:zhengying@fudan.edu.cn)

## Abstract

**Background:** Over the recent years, machine learning methods have been increasingly explored in cancer prognosis because of the appearance of improved machine learning algorithms. These algorithms can use censored data for modeling, such as support vector machines for survival analysis and random survival forest (RSF). However, it is still debated whether traditional (Cox proportional hazard regression) or machine learning-based prognostic models have better predictive performance.

**Objective:** This study aimed to compare the performance of breast cancer prognostic prediction models based on machine learning and Cox regression.

**Methods:** This retrospective cohort study included all patients diagnosed with breast cancer and subsequently hospitalized in Fudan University Shanghai Cancer Center between January 1, 2008, and December 31, 2016. After all exclusions, a total of 22,176 cases with 21 features were eligible for model development. The data set was randomly split into a training set (15,523 cases, 70%) and a test set (6653 cases, 30%) for developing 4 models and predicting the overall survival of patients diagnosed with breast cancer. The discriminative ability of models was evaluated by the concordance index (C-index), the time-dependent area under the curve, and D-index; the calibration ability of models was evaluated by the Brier score.

**Results:** The RSF model revealed the best discriminative performance among the 4 models with 3-year, 5-year, and 10-year time-dependent area under the curve of 0.857, 0.838, and 0.781, a D-index of 7.643 (95% CI 6.542, 8.930) and a C-index of 0.827 (95% CI 0.809, 0.845). The statistical difference of the C-index was tested, and the RSF model significantly outperformed the Cox-EN (elastic net) model (C-index 0.816, 95% CI 0.796, 0.836;  $P=.01$ ), the Cox model (C-index 0.814, 95% CI 0.794, 0.835;  $P=.003$ ), and the support vector machine model (C-index 0.812, 95% CI 0.793, 0.832;  $P<.001$ ). The 4 models' 3-year, 5-year, and 10-year Brier scores were very close, ranging from 0.027 to 0.094 and less than 0.1, which meant all models had good calibration. In the context of feature importance, elastic net and RSF both indicated that TNM staging, neoadjuvant therapy, number of lymph node metastases, age, and tumor diameter were the top 5 important features for predicting the prognosis of breast cancer. A final online tool was developed to predict the overall survival of patients with breast cancer.

**Conclusions:** The RSF model slightly outperformed the other models on discriminative ability, revealing the potential of the RSF method as an effective approach to building prognostic prediction models in the context of survival analysis.

(*JMIR Med Inform* 2022;10(2):e33440) doi:[10.2196/33440](https://doi.org/10.2196/33440)

**KEYWORDS**

breast cancer; machine learning; survival analysis; random survival forest; support vector machine; medical informatics; prediction models

## Introduction

Breast cancer is a leading cause of morbidity and mortality in women worldwide, and the prediction of breast cancer prognosis is crucial for decision-making. Accurate outcome prediction can assist doctors with providing appropriate treatment plans for patients, which in turn could improve their chances of survival and lessen the suffering. Several prognostic prediction models have already been developed. PREDICT and Adjuvant! Online are 2 famous prognostic prediction tools for breast cancer based on clinical and pathological characteristics [1,2]. These models have been validated by external data set and are commonly used in the United States and Western Europe. However, several external validations that were made in Asian countries revealed a less-than-optimal predictive ability [3-6].

For survival analysis of follow-up observations, the most important challenge is dealing with censored data. The Cox proportional hazard regression is a classical modeling method used to analyze right-censored data in survival analysis with good interpretability. Typically, the Cox proportional hazard regression imposes proportional hazard assumption and the assumption that continuous covariates have a linear effect on the logarithm of the hazard, which the real-world data may not satisfy [7]. Compared with the Cox proportional hazard regression, machine learning methods do not make any parametric or semiparametric assumptions and have the ability to detect and account for higher-order interactions as well as nonlinear relationships [8]. While there have been some attempts to use machine learning to build cancer prognosis prediction models [6,9-13], currently, there is no consensus on whether traditional or machine learning-based prognostic prediction models have a better predictive performance.

Here, we discuss two main types of prognostic prediction models using machine learning algorithms. The first types are the binary classification models, which give a probability of the interested outcome at a specific time. Several studies have used machine learning methods to generate prognostic prediction models based on classification. The outcome variable of these models is the status of survival at 5 years [14-17] or at the time of data collection [18,19]. The limitation of these models is that they are not able to include right-censored observations that were censored before the specified time, because the outcome of these observations is unknown. Moreover, using the classification outcome (survival status at a specific time) instead of the survival outcome (survival time and status of the censor) can lead to a loss of information. The second types are models using improved algorithms of original machine learning algorithms to enable modeling and analysis of censored data, such as support vector machines (SVM) for survival analysis [20] and random survival forest (RSF) [21]. These methods can describe probability (RSF) and risk scores (SVM and RSF) of the interested outcomes at different time points rather than at a specific time point and can consider both the survival time and the status of the censor.

In this study, traditional (Cox) and machine learning-based (SVM and RSF) prognostic prediction models were developed for patients with breast cancer based on a large cohort of Chinese patients diagnosed with breast cancer and hospitalized in Fudan University Shanghai Cancer Center. We aimed to compare the performance of different models to pick the optimal predictive model and provide a reference for the development of machine learning in the prognosis prediction of breast cancer.

## Methods

### Study Design and Ethical Considerations

This retrospective cohort study included all patients diagnosed with breast cancer and subsequently hospitalized in Fudan University Shanghai Cancer Center between January 1, 2008, and December 31, 2016. Data containing demographic and clinicopathologic features were obtained from the hospital information system. Overall Survival, defined as the duration between the time of first treatment and the date of death, was taken as the outcome to build the predictive models. The outcome information was derived from medical visit records, telephone visits, and death certificate data linkage with the cancer registry system or death certificate system run by the provincial Centers for Disease Control and Prevention.

By March 1, 2021, medical information and follow-up information were collected from 25,629 patients. After excluding male patients, patients with bilateral breast cancer (362 cases), and patients with  $\geq 3$  missing features, 22,176 cases with 21 features were eligible for further analysis. Patients were followed for a median follow-up time of 68.9 months (95% CI 68.42, 69.33). The data set was then randomly split into a training set (15,523 cases, 70%) and a test set (6653 cases, 30%). The statistical description of features and the survival curves of patients in the training and test set are shown in Table S1 and Figure S1 in [Multimedia Appendix 1](#).

This study was approved by the Fudan University Shanghai Cancer Center Institutional Review Board (Registration YF-2021-01).

### Preprocess of Missing Data

Since the data were generated and collected in a real medical environment, there were many observations with missing features. As the SVM and RSF methods do not support the analysis of data sets with missing values, we performed a 2-step process in order to reduce the impact of missing values on the training process of developing prediction models. Firstly, we excluded patients with too many missing features. The number of missing features of patients and the log-rank test results are shown in Table S1 in [Multimedia Appendix 2](#). The log-rank method was used to test the difference between the survival state of 25,267 patients and the remaining patients. Based on the results of the log-rank test, when we excluded patients with  $\geq 3$  missing features, there was no significant difference between the survival of the remaining patients (22,176 cases) and the

survival of the overall patients (25,267 cases;  $P=.17$ ). Therefore, 3 was taken as the cut-off value, and patients with  $\geq 3$  missing features were excluded. The statistics for missing features before and after the first step of processing are shown in Table S2 in [Multimedia Appendix 2](#), and the remaining 22,176 cases are eligible for further analysis. Secondly, the remaining missing data were imputed by the missForest algorithm using library “missingpy” (0.2.0) in Python (Python Software Foundation). MissForest is a nonparametric imputation method that could be applied for both continuous and categorical variables and does not make explicit assumptions about the functional form of the data [22]. In the process of imputing the missing values, the outcome data were not involved in case imputed data were affected and falsely related to the outcome data.

### Statistical Analysis

The objective outcome in the study was time to event, which is right-censored survival data. Therefore, the following 3 survival modeling approaches were used to predict the survival time of patients diagnosed with breast cancer: Cox proportional hazard regression [23], SVM [24], and RSF [21]. Elastic Net (EN) was used as the feature selection method to screen important features to train the 3 models. Technical implementation details, including the libraries and the process of hyperparameter tuning, are provided in the [Multimedia Appendix 3](#). Moreover, we have open sourced the Python and R code that we developed for generating the models and evaluating the performance of the models in the GitHub repository [25].

The Cox proportional hazard regression is a classical modeling method for survival analysis. The model predicts the probability that the event of interest has occurred at a given time for given values of the predictor variables [23]. We added a traditional feature selection method for the Cox model, where univariate Cox analysis was performed before significant ( $P<.1$ ) and clinically relevant features were forced into multivariate Cox regression analysis. The Cox model using the EN method was named “Cox-EN,” and the one using the traditional variable selection method was named “Cox.”

Usually, the predictors should satisfy the proportional hazard assumption in the Cox model. However, the main goal of modeling in this study was survival prediction and maximizing concordance index (C-index) and time-dependent area under the curve (AUC), regardless of how predictions are generated. Therefore, we did not perform the test for proportional hazards in the process of modeling [26].

SVM is a supervised machine learning algorithm, which can be used for both classification and regression challenges. An SVM model is basically a representation of different classes in a hyperplane in multidimensional space. The hyperplane is generated iteratively by SVM so that the error can be minimized. The goal of SVM is to divide the data sets into classes to find a maximum marginal hyperplane [24].

Several extensions of SVM to survival analysis were proposed. Shivaswamy et al [27] introduced an approach for censored targets by casting survival analysis as a regression problem. Van Belle et al [24,28] proposed the ranking approach and the hybrid approach combining the regression and ranking approach

for survival outcomes. As an objective function of the ranking-based technique depends on a quadratic number of constraints with respect to the number of training samples, which makes training intractable with medium to large-sized data sets, we chose an approach of efficient training of linear survival SVM [20].

RSF, which was developed by Ishwaran et al [21], is an ensemble of tree-based learners for survival analysis of right-censored data and an extension of the random forest method. Using independent bootstrap samples, each tree in RSF is grown by randomly selecting a subset of features for each node and then splitting the node using a survival criterion involving information of survival time and censoring status [21].

EN is a feature screening technique that uses the penalties L1 and L2 from both the least absolute shrinkage and selection operator (LASSO) and ridge techniques to regularize regression models. The EN method is improved based on the shortcomings of both ridge and LASSO methods. The ridge method keeps all the features and cannot perform the function of feature screening. When it comes to multiple correlated features, the LASSO method randomly picks one of these features from such groups and entirely ignores the rest, while the EN method is likely to pick a few at once [29].

### Evaluation of Model Performance

The discriminative ability of models was evaluated by the C-index [30], time-dependent AUC [31], and D-index [32]. C-index measures the overall discriminative ability of models, while time-dependent AUC measures the discriminative ability of models by comparing the predicted probabilities with the actual binary survival status and the probability estimation of a death outcome of censored observations at an interested time. C-index and time-dependent AUC both range in an interval from 0 to 1, and a value of 0.5 is comparable to random guessing, while a value of 1 means perfect discrimination. D-index was used to measure the separation between patients from equally sized high-risk and low-risk groups divided according to the risk score obtained from different models. Higher values of D-index indicate a more remarkable discriminative ability of the model. The survival curves of high-risk and low-risk groups was estimated using the Kaplan-Meier method, and the log-rank test was used to compare survival curves. The calibration ability of models was evaluated by the Brier score [33], which varies between 0 and 1, while a lower Brier score was indicative of a better-calibrated prediction. A value of 0.25 is comparable to random guessing, while a value of 0 means perfect discrimination.

## Results

### User and Model Statistics

A total of 22,176 patients with 68.9 months (95% CI 68.42, 69.33) of median follow-up were included in this study. We fitted 4 prognostic models (Cox, Cox-EN, RSF, and SVM) for predicting the overall survival of breast cancer patients with the training set and then used C-index, time-dependent AUC, D-index, and Brier score to evaluate them in the independent

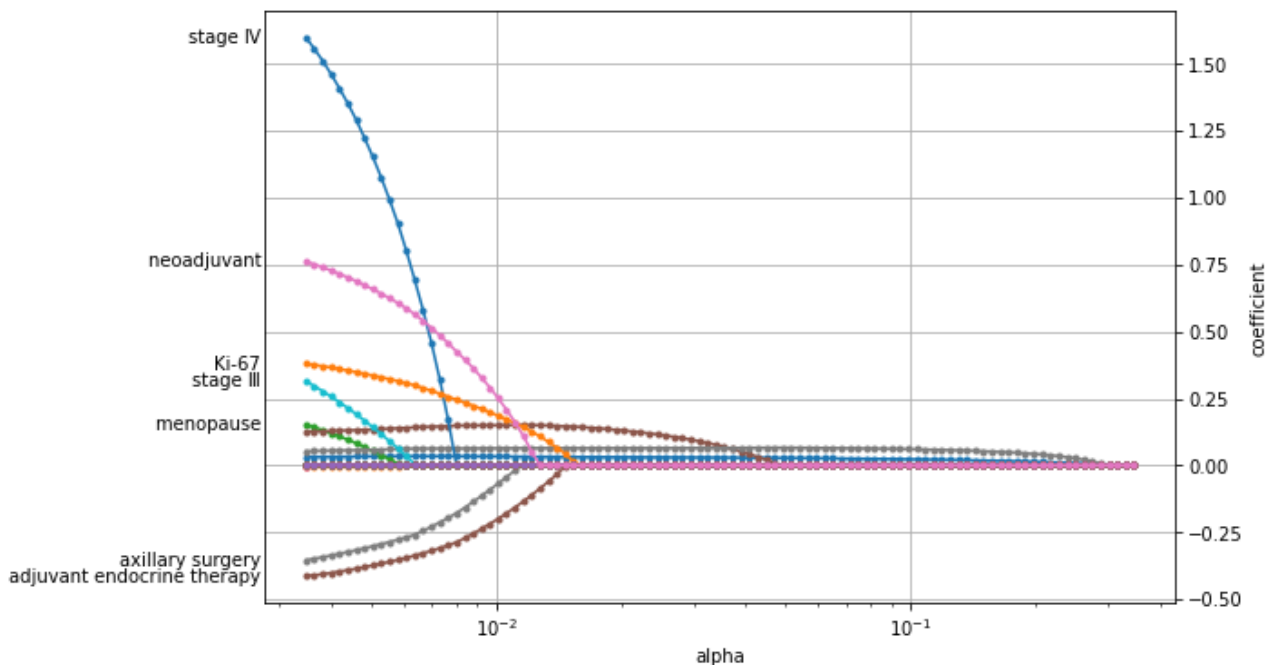
test set. All models showed good calibration, and RSF outperformed other models on discriminative ability with a C-index of 0.827 (95% CI 0.809, 0.845).

### Evaluation of Feature Importance

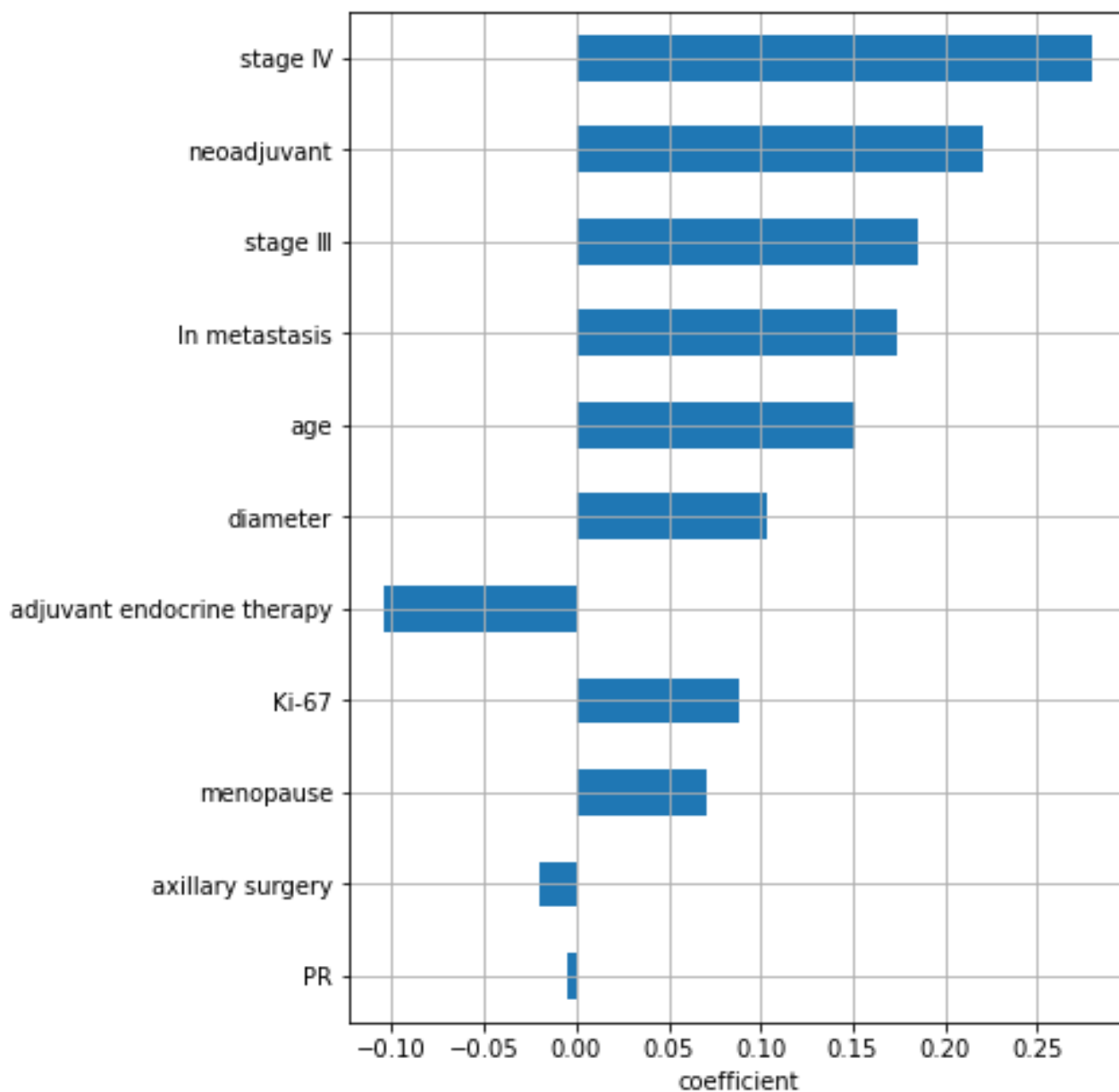
In order to screen out features with a large contribution to predicting the prognosis of breast cancer, the EN was first used to select important features, resulting in a total of 21 features. The ways the coefficients changed for varying  $\alpha$  is shown in Figure 1, and the coefficient of each feature corresponding to the optimal  $\alpha$  is shown in Figure 2. The top 5 important features were TNM staging, neoadjuvant therapy, number of lymph node metastases, age, and diameter of the tumor. RSF was used to rank the importance of the 11 features selected by the EN, and the results are shown in Figure 3. The top 5 important features were the number of lymph node metastasis, age, tumor diameter, neoadjuvant therapy, and TNM staging.

The results of univariate and multivariate Cox analysis are shown in Multimedia Appendix 4. Except for cases of the side of the tumor, multiple tumors, adjuvant chemotherapy, and targeted therapy, all features had a  $P$  value of less than .1 in the univariate analysis. Considering that adjuvant chemotherapy and targeted therapy could be confounding factors, multivariate analysis was performed using adjuvant chemotherapy, targeted therapy, and the significant factors ( $P < .1$ ) from univariate analysis. The results of the multivariate analyses showed that age, menopause, invasive, diameter, lymph node metastasis, TNM, Ki 67, estrogen receptors, progesterone receptors, breast surgery, axillary surgery, adjuvant chemotherapy, targeted therapy, adjuvant radiotherapy, adjuvant endocrine therapy, and neoadjuvant therapy had a  $P$  value of less than .05, and the Cox model was built by these features.

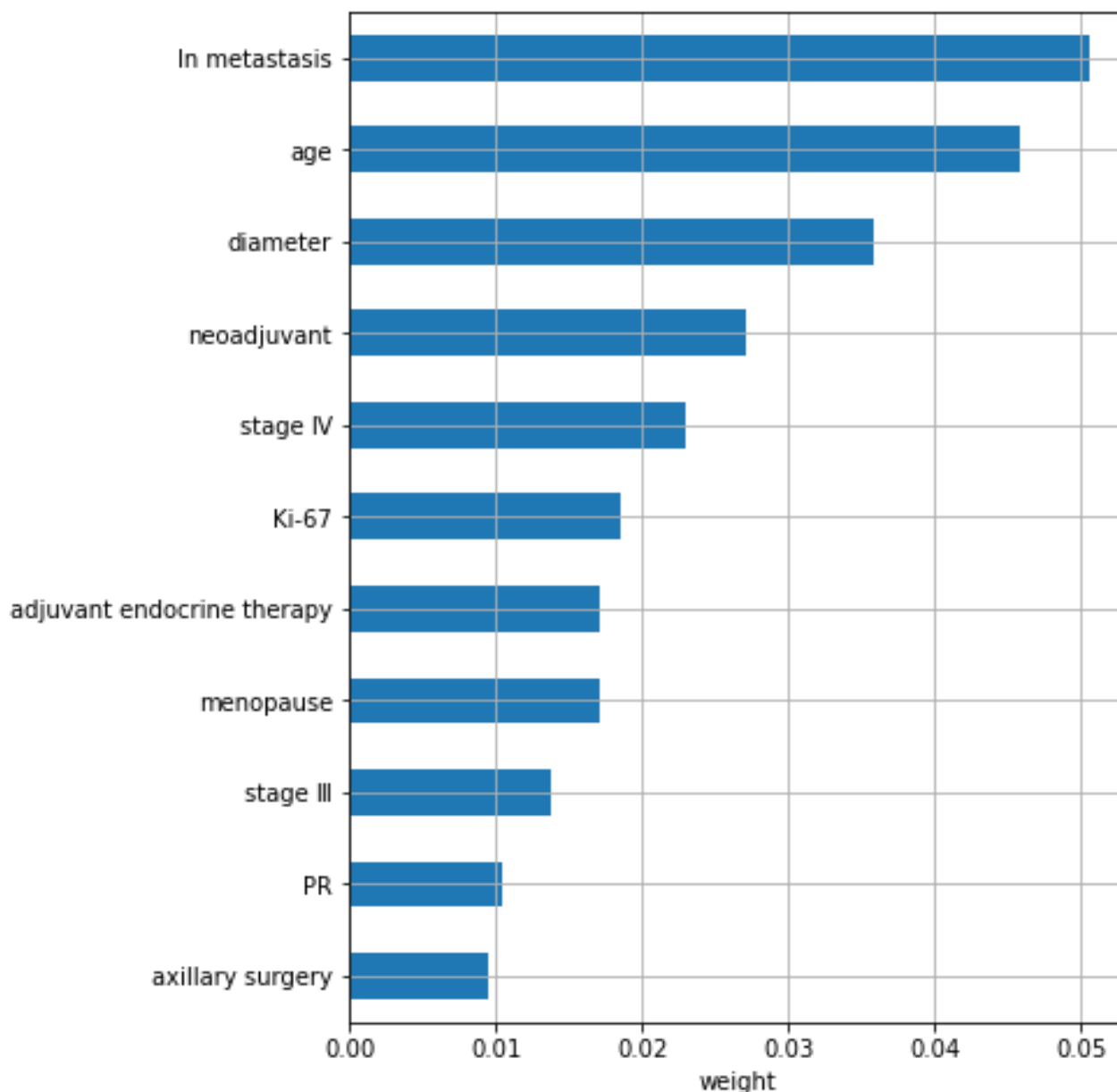
**Figure 1.** The coefficients of features change for varying  $\alpha$ .



**Figure 2.** The important coefficient of each feature corresponding to the optimal  $\alpha$  by elastic net. Ln: lymph node; PR: progesterone receptors.



**Figure 3.** The important coefficient of each feature by random survival forest. Ln: lymph node; PR: progesterone receptors.



### Methods Performance

Evaluation results of the 4 models are shown in Table 1. From the point of view of the C-index, the RSF model slightly and significantly outperformed the Cox-EN model ( $P=.01$ ), the Cox model ( $P=.003$ ), and the SVM model ( $P<.001$ ) on discriminative ability, and no significant difference was found between the discriminative ability of other models. Time-dependent receiver operating characteristic curves of each model at 3 years, 5 years, and 10 years are shown in Figure 4. The time-dependent AUC

of each model over time is shown in Figure 5. As shown in Figure 5, the time-dependent AUC of RSF was the highest at most times. Survival curves of the high-risk and low-risk groups divided according to the risk score are shown in Figure 6. The D-index of 7.643 from the RSF model was also the highest, and it can be interpreted as the risk of death in the high-risk group, which is 7.643 times the risk of death in the low-risk group. The 4 models' 3-year, 5-year, and 10-year Brier scores were all  $<0.1$ , suggesting that all models had good calibration.



**Table 1.** Performance of different methods.

Indexes	Cox	Cox-EN <sup>a</sup>	SVM <sup>b</sup>	RSF <sup>c</sup>
C-index <sup>d</sup> (95% CI)	0.814 (0.794,0.835)	0.816 (0.796,0.836)	0.812 (0.793,0.832)	0.827 (0.809,0.845)
AUC <sup>e</sup> (3 years)	0.850	0.857	0.847	0.857
AUC (5 years)	0.821	0.822	0.823	0.838
AUC (10 years)	0.770	0.769	0.760	0.781
D-index (95% CI)	7.210 (6.172,8.424)	7.466 (6.383,8.733)	6.522 (5.606,7.583)	7.643 (6.542,8.930)
Brier score (3 years)	0.027	0.027	— <sup>f</sup>	0.027
Brier score (5 years)	0.044	0.045	—	0.045
Brier score (10 years)	0.094	0.093	—	0.093

<sup>a</sup>EN: elastic net.

<sup>b</sup>SVM: support vector machine.

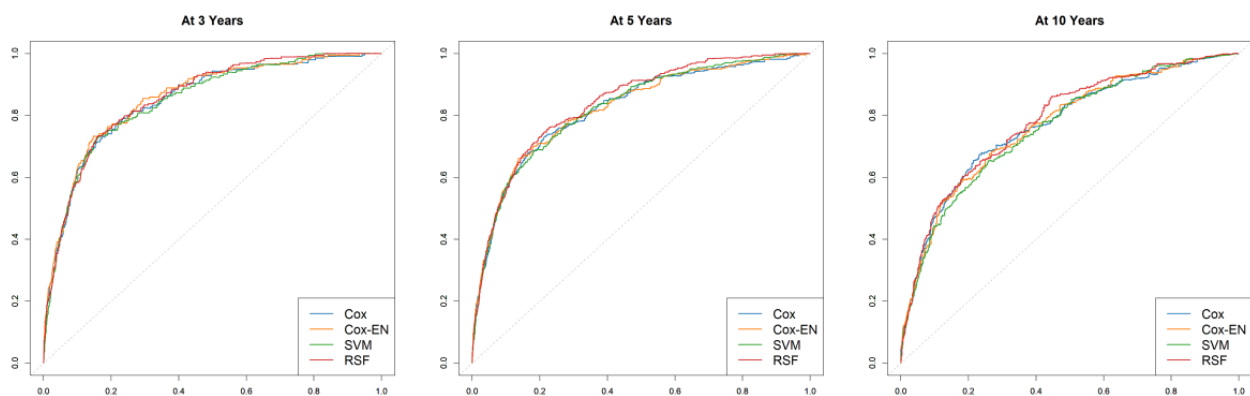
<sup>c</sup>RSF: random survival forest.

<sup>d</sup>C-index: concordance index.

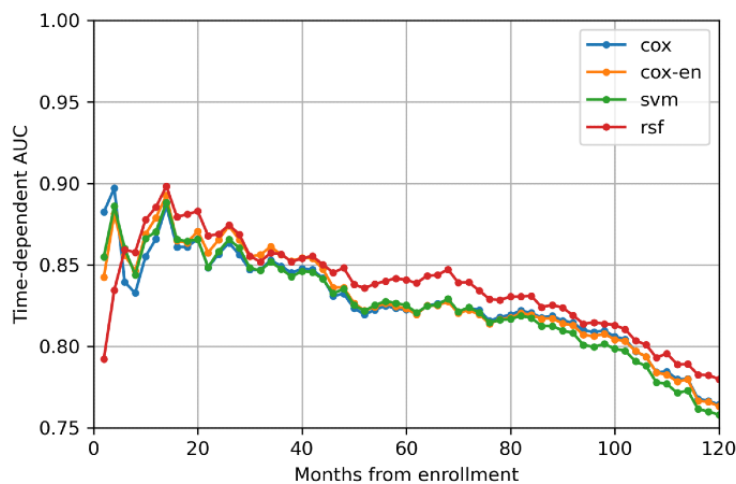
<sup>e</sup>AUC: area under the curve.

<sup>f</sup>Not available. Survival support vector machine can only predict a risk score and not a probability. Therefore, Brier score is not available for survival support vector machine.

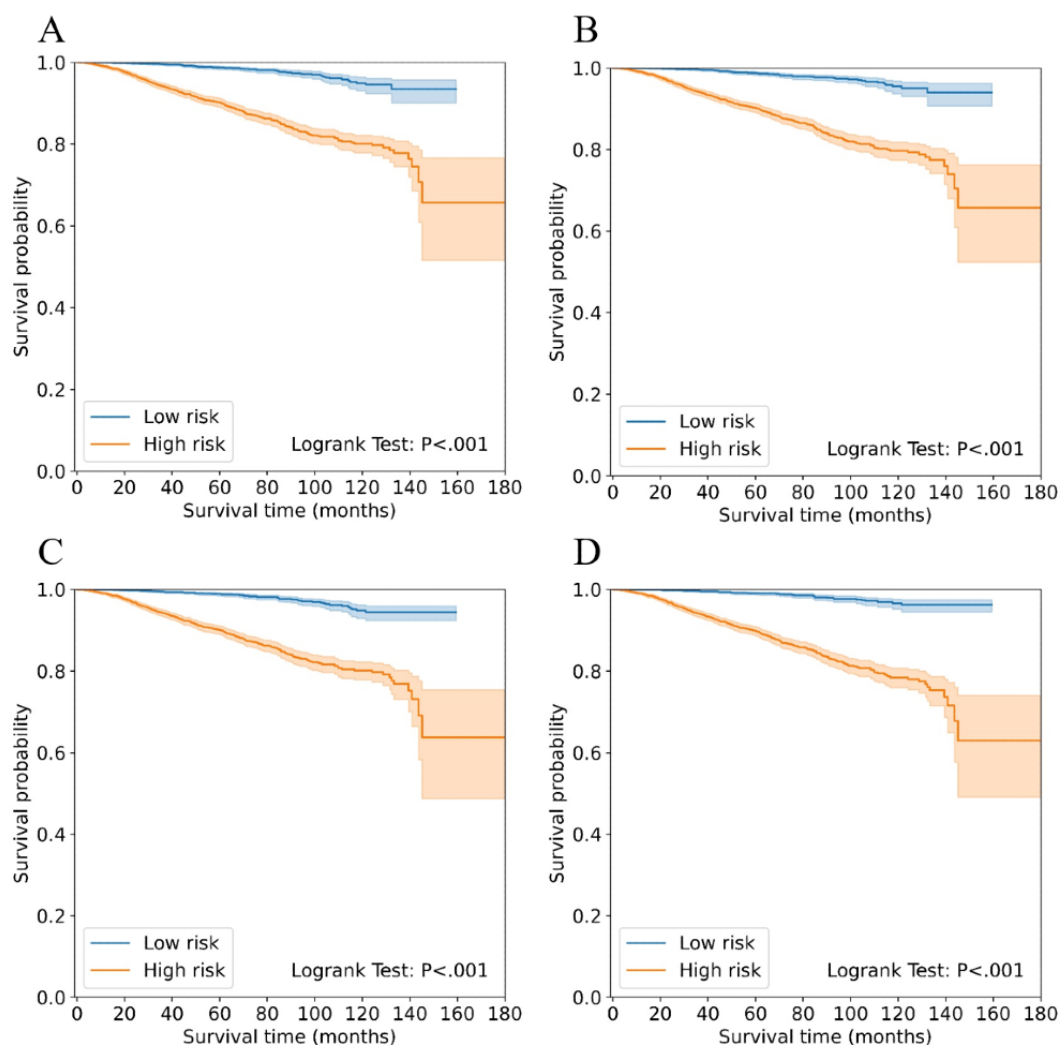
**Figure 4.** Time-dependent receiver operating characteristic curves of models at 3 years, 5 years, and 10 years. EN: elastic net; RSF: random survival forest; SVM: support vector machine.



**Figure 5.** Time-dependent AUC of models over time. AUC: area under the curve; EN: elastic net; RSF: random survival forest; SVM: support vector machine.



**Figure 6.** Survival curves of high-risk and low-risk groups divided according to the risk score from (A) Cox, (B) Cox-EN (elastic net), (C) SVM (support vector machine), and (D) RSF (random survival forest).



### Online Prognostic Prediction Tool

Although the RSF model achieved the best performance among these models, the interpretability and computational efficiency of the RSF model had to be considered at the same time in the deployment of the online prognostic prediction tool. The memory usage of the RSF model was too large for the model to be deployed on a website and have good computational efficiency. The Cox-EN model achieved suboptimal performance in the study and had better interpretability and computational efficiency compared with the RSF model. Therefore, it was selected as a backend for the online prognostic prediction tool [34].

### Discussion

In this paper, we compared the performance of traditional (Cox) and machine learning-based (SVM and RSF) prognostic prediction models for patients diagnosed with breast cancer and found out the RSF model slightly and significantly outperformed the Cox-EN model, the Cox model, and the SVM model on discriminative ability. Compared with Cox, Cox-EN, and SVM, the RSF model had a slightly better performance with a C-index

of 0.827 (95% CI 0.809, 0.845) and 3-year, 5-year, and 10-year time-dependent AUC of 0.857, 0.838, and 0.781, respectively. The results in this study were similar to those reported by some previous studies. For example, Liu et al [10] used several methods, including RSF and Cox, to predict breast cancer progression with a sample size of 4575 patients. The results showed that the RSF model achieved better performance with a C-index of 0.814 compared with the Cox model with a C-index of 0.759. Rahman et al [35] showed that RSF (5-year time-dependent AUC 0.839, 95% CI 0.826, 0.849) outperformed Cox (5-year time-dependent AUC 0.823, 95% CI 0.811, 0.833) in the survival prediction of patients with esophageal cancer.

The possible reason for RSF achieving better performance may be that RSF is able to detect and account for higher-order interactions and nonlinear relationships. However, despite the great predictive performance of RSF, there are several shortcomings that limit the wide adoption of RSF. Firstly, the theoretical properties and the inferential procedures of RSF are not well understood. Secondly, RSF creates a “black-box” model that is hard to interpret or visualize [8]. Nonetheless, RSF still has the potential to be used as an effective approach to build prognostic prediction models in the context of survival analysis.

A major advantage of this paper was the large-scale prospective cohort design with a long follow-up time. To the best of our knowledge, this is the study with the largest sample size for breast cancer prognostic prediction modeling based on machine learning in the Chinese population thus far. Even though the study is based on a single institution, the large-scale prospective cohort and long follow-up time make the results valuable and credible.

There are some limitations in this study that should be acknowledged. The main limitation is that this study was performed in a single center in China with no external validation. Therefore, the current results need further multi-institutional validation with larger samples before the prediction models could be used in clinical practice. Another limitation relates to missing data that were imputed, and we could not ascertain the

effect of the imputation of missing data on the overall results and subsequent conclusions. Moreover, we chose the randomized search method with 50 parameter settings sampled instead of grid search in the process of tuning the hyperparameters of the RSF due to the limitation of the computational efficiency. This may cause an underestimate of the performance of the RSF model.

In summary, the RSF model slightly outperformed the other models on discriminative ability, revealing the potential of the RSF method to be used as an effective approach to build prognostic prediction models in the context of survival analysis. Our future work will focus on additional external validation of the model using data from multiple centers to verify the extrapolation of our results.

---

## Conflicts of Interest

None declared.

---

### Multimedia Appendix 1

The statistical description of features and the survival curves of patients in the training and test set.

[[DOCX File , 204 KB - medinform\\_v10i2e33440\\_app1.docx](#) ]

---

### Multimedia Appendix 2

Statistics for missing fields and missing features before and after processing.

[[DOCX File , 16 KB - medinform\\_v10i2e33440\\_app2.docx](#) ]

---

### Multimedia Appendix 3

Technical implementation details.

[[DOCX File , 16 KB - medinform\\_v10i2e33440\\_app3.docx](#) ]

---

### Multimedia Appendix 4

Results of univariate survival analysis and multivariate survival analysis.

[[DOCX File , 18 KB - medinform\\_v10i2e33440\\_app4.docx](#) ]

---

## References

1. Candido Dos Reis FJ, Wishart GC, Dicks EM, Greenberg D, Rashbass J, Schmidt MK, et al. An updated PREDICT breast cancer prognostication and treatment benefit prediction model with independent validation. *Breast Cancer Res* 2017 May 22;19(1):58 [[FREE Full text](#)] [doi: [10.1186/s13058-017-0852-3](https://doi.org/10.1186/s13058-017-0852-3)] [Medline: [28532503](https://pubmed.ncbi.nlm.nih.gov/28532503/)]
2. Ravdin PM, Siminoff LA, Davis GJ, Mercer MB, Hewlett J, Gerson N, et al. Computer program to assist in making decisions about adjuvant therapy for women with early breast cancer. *J Clin Oncol* 2001 Feb 15;19(4):980-991. [doi: [10.1200/JCO.2001.19.4.980](https://doi.org/10.1200/JCO.2001.19.4.980)] [Medline: [11181660](https://pubmed.ncbi.nlm.nih.gov/11181660/)]
3. Bhoo-Pathy N, Yip C, Hartman M, Saxena N, Taib NA, Ho G, et al. Adjuvant! Online is overoptimistic in predicting survival of Asian breast cancer patients. *Eur J Cancer* 2012 May;48(7):982-989 [[FREE Full text](#)] [doi: [10.1016/j.ejca.2012.01.034](https://doi.org/10.1016/j.ejca.2012.01.034)] [Medline: [22366561](https://pubmed.ncbi.nlm.nih.gov/22366561/)]
4. Wong H, Subramaniam S, Alias Z, Taib NA, Ho G, Ng C, et al. The predictive accuracy of PREDICT: a personalized decision-making tool for Southeast Asian women with breast cancer. *Medicine (Baltimore)* 2015 Feb;94(8):e593 [[FREE Full text](#)] [doi: [10.1097/MD.0000000000000593](https://doi.org/10.1097/MD.0000000000000593)] [Medline: [25715267](https://pubmed.ncbi.nlm.nih.gov/25715267/)]
5. Zaguirre K, Kai M, Kubo M, Yamada M, Kurata K, Kawaji H, et al. Validity of the prognostication tool PREDICT version 2.2 in Japanese breast cancer patients. *Cancer Med* 2021 Mar;10(5):1605-1613 [[FREE Full text](#)] [doi: [10.1002/cam4.3713](https://doi.org/10.1002/cam4.3713)] [Medline: [33452761](https://pubmed.ncbi.nlm.nih.gov/33452761/)]
6. Zhong X, Luo T, Deng L, Liu P, Hu K, Lu D, et al. Multidimensional Machine Learning Personalized Prognostic Model in an Early Invasive Breast Cancer Population-Based Cohort in China: Algorithm Validation Study. *JMIR Med Inform* 2020 Nov 09;8(11):e19069 [[FREE Full text](#)] [doi: [10.2196/19069](https://doi.org/10.2196/19069)] [Medline: [33164899](https://pubmed.ncbi.nlm.nih.gov/33164899/)]

7. Goerdten J, Carrière I, Muniz-Terrera G. Comparison of Cox proportional hazards regression and generalized Cox regression models applied in dementia risk prediction. *Alzheimers Dement (N Y)* 2020;6(1):e12041 [[FREE Full text](#)] [doi: [10.1002/trc2.12041](https://doi.org/10.1002/trc2.12041)] [Medline: [32548239](#)]
8. Hu C, Steingrimsson JA. Personalized Risk Prediction in Clinical Oncology Research: Applications and Practical Issues Using Survival Trees and Random Forests. *J Biopharm Stat* 2018;28(2):333-349 [[FREE Full text](#)] [doi: [10.1080/10543406.2017.1377730](https://doi.org/10.1080/10543406.2017.1377730)] [Medline: [29048993](#)]
9. Senders JT, Staples P, Mehrtash A, Cote DJ, Taphoorn MJB, Reardon DA, et al. An Online Calculator for the Prediction of Survival in Glioblastoma Patients Using Classical Statistics and Machine Learning. *Neurosurgery* 2020 Feb 01;86(2):E184-E192 [[FREE Full text](#)] [doi: [10.1093/neuros/nyz403](https://doi.org/10.1093/neuros/nyz403)] [Medline: [31586211](#)]
10. Liu P, Fu B, Yang SX, Deng L, Zhong X, Zheng H. Optimizing Survival Analysis of XGBoost for Ties to Predict Disease Progression of Breast Cancer. *IEEE Trans Biomed Eng* 2021 Jan;68(1):148-160. [doi: [10.1109/TBME.2020.2993278](https://doi.org/10.1109/TBME.2020.2993278)] [Medline: [32406821](#)]
11. Qiu X, Gao J, Yang J, Hu J, Hu W, Kong L, et al. A Comparison Study of Machine Learning (Random Survival Forest) and Classic Statistic (Cox Proportional Hazards) for Predicting Progression in High-Grade Glioma after Proton and Carbon Ion Radiotherapy. *Front Oncol* 2020;10:551420 [[FREE Full text](#)] [doi: [10.3389/fonc.2020.551420](https://doi.org/10.3389/fonc.2020.551420)] [Medline: [33194609](#)]
12. Tran BX, Latkin CA, Sharafeldin N, Nguyen K, Vu GT, Tam WWS, et al. Characterizing Artificial Intelligence Applications in Cancer Research: A Latent Dirichlet Allocation Analysis. *JMIR Med Inform* 2019 Sep 15;7(4):e14401 [[FREE Full text](#)] [doi: [10.2196/14401](https://doi.org/10.2196/14401)] [Medline: [31573929](#)]
13. Cos H, Li D, Williams G, Chininis J, Dai R, Zhang J, et al. Predicting Outcomes in Patients Undergoing Pancreatectomy Using Wearable Technology and Machine Learning: Prospective Cohort Study. *J Med Internet Res* 2021 Mar 18;23(3):e23595 [[FREE Full text](#)] [doi: [10.2196/23595](https://doi.org/10.2196/23595)] [Medline: [33734096](#)]
14. Abdikenov B, Iklassov Z, Sharipov A, Hussain S, Jamwal PK. Analytics of Heterogeneous Breast Cancer Data Using Neuroevolution. *IEEE Access* 2019;7:18050-18060. [doi: [10.1109/access.2019.2897078](https://doi.org/10.1109/access.2019.2897078)]
15. Zhao M, Tang Y, Kim H, Hasegawa K. Machine Learning With K-Means Dimensional Reduction for Predicting Survival Outcomes in Patients With Breast Cancer. *Cancer Inform* 2018;17:1176935118810215 [[FREE Full text](#)] [doi: [10.1177/1176935118810215](https://doi.org/10.1177/1176935118810215)] [Medline: [30455569](#)]
16. García-Laencina PJ, Abreu PH, Abreu MH, Afonso N. Missing data imputation on the 5-year survival prediction of breast cancer patients with unknown discrete values. *Comput Biol Med* 2015 Apr;59:125-133. [doi: [10.1016/j.compbiomed.2015.02.006](https://doi.org/10.1016/j.compbiomed.2015.02.006)] [Medline: [25725446](#)]
17. Lotfnezhad Afshar H, Ahmadi M, Roudbari M, Sadoughi F. Prediction of breast cancer survival through knowledge discovery in databases. *Glob J Health Sci* 2015 Jan 26;7(4):392-398 [[FREE Full text](#)] [doi: [10.5539/gjhs.v7n4p392](https://doi.org/10.5539/gjhs.v7n4p392)] [Medline: [25946945](#)]
18. Chao C, Yu Y, Cheng B, Kuo Y. Construction the model on the breast cancer survival analysis use support vector machine, logistic regression and decision tree. *J Med Syst* 2014 Oct;38(10):106. [doi: [10.1007/s10916-014-0106-1](https://doi.org/10.1007/s10916-014-0106-1)] [Medline: [25119239](#)]
19. Ganggayah MD, Taib NA, Har YC, Lio P, Dhillon SK. Predicting factors for survival of breast cancer patients using machine learning techniques. *BMC Med Inform Decis Mak* 2019 Mar 22;19(1):48 [[FREE Full text](#)] [doi: [10.1186/s12911-019-0801-4](https://doi.org/10.1186/s12911-019-0801-4)] [Medline: [30902088](#)]
20. Pölsterl S, Amin Katouzian NN. Fast Training of Support Vector Machines for Survival Analysis. 2015 Presented at: ECML PKDD: Joint European Conference on Machine Learning and Knowledge Discovery in Databases; September 7-11, 2015; Porto, Portugal p. 243-259 URL: [https://link.springer.com/chapter/10.1007/978-3-319-23525-7\\_15](https://link.springer.com/chapter/10.1007/978-3-319-23525-7_15) [doi: [10.1007/978-3-319-23525-7\\_15](https://doi.org/10.1007/978-3-319-23525-7_15)]
21. 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](https://doi.org/10.1214/08-AOAS169)]
22. Stekhoven DJ, Bühlmann P. MissForest--non-parametric missing value imputation for mixed-type data. *Bioinformatics* 2012 Jan 01;28(1):112-118. [doi: [10.1093/bioinformatics/btr597](https://doi.org/10.1093/bioinformatics/btr597)] [Medline: [22039212](#)]
23. Holford TR. Life tables with concomitant information. *Biometrics* 1976 Sep;32(3):587-597. [Medline: [963172](#)]
24. Van Belle V, Kristiaan P, Suykens J, Van Huffel S. Support vector machines for survival analysis. 2007 Jan 01 Presented at: Third International Conference on Computational Intelligence in Medicine and Healthcare (CIMED2007); July 1-7, 2007; Plymouth, England URL: [https://limo.libis.be/primo-explore/fulldisplay?docid=LIRIAS1690710&context=L&vid=Lirias&search\\_scope=Lirias&tab=default\\_tab&lang=en\\_US&fromSitemap=1](https://limo.libis.be/primo-explore/fulldisplay?docid=LIRIAS1690710&context=L&vid=Lirias&search_scope=Lirias&tab=default_tab&lang=en_US&fromSitemap=1)
25. ml-for-survival. GitHub. URL: <https://github.com/xiaojialong0518/ml-for-survival> [accessed 2021-12-12]
26. Stensrud MJ, Hernán MA. Why Test for Proportional Hazards? *JAMA* 2020 Apr 14;323(14):1401-1402. [doi: [10.1001/jama.2020.1267](https://doi.org/10.1001/jama.2020.1267)] [Medline: [32167523](#)]
27. Shivaswamy PK, Chu W, Jansche M. A support vector approach to censored targets. 2007 Oct 28 Presented at: Seventh IEEE International Conference on Data Mining; October 28-31, 2007; Omaha, NE, USA. [doi: [10.1109/icdm.2007.93](https://doi.org/10.1109/icdm.2007.93)]
28. Van Belle V, Pelckmans K, Van Huffel S, Suykens JAK. Support vector methods for survival analysis: a comparison between ranking and regression approaches. *Artif Intell Med* 2011 Oct;53(2):107-118. [doi: [10.1016/j.artmed.2011.06.006](https://doi.org/10.1016/j.artmed.2011.06.006)] [Medline: [21821401](#)]

29. Zou H, Hastie T. Regularization and variable selection via the elastic net. *J Royal Statistical Soc B* 2005 Apr;67(2):301-320. [doi: [10.1111/j.1467-9868.2005.00503.x](https://doi.org/10.1111/j.1467-9868.2005.00503.x)]
30. Uno H, Cai T, Pencina MJ, D'Agostino RB, Wei LJ. On the C-statistics for evaluating overall adequacy of risk prediction procedures with censored survival data. *Stat Med* 2011 May 10;30(10):1105-1117 [FREE Full text] [doi: [10.1002/sim.4154](https://doi.org/10.1002/sim.4154)] [Medline: [21484848](https://pubmed.ncbi.nlm.nih.gov/21484848/)]
31. Heagerty PJ, Lumley T, Pepe MS. Time-dependent ROC curves for censored survival data and a diagnostic marker. *Biometrics* 2000 Jun;56(2):337-344. [doi: [10.1111/j.0006-341x.2000.00337.x](https://doi.org/10.1111/j.0006-341x.2000.00337.x)] [Medline: [10877287](https://pubmed.ncbi.nlm.nih.gov/10877287/)]
32. Royston P, Sauerbrei W. A new measure of prognostic separation in survival data. *Stat Med* 2004 Mar 15;23(5):723-748. [doi: [10.1002/sim.1621](https://doi.org/10.1002/sim.1621)] [Medline: [14981672](https://pubmed.ncbi.nlm.nih.gov/14981672/)]
33. Brier GW. Verification of forecasts expressed in terms of probability. *Mon. Wea. Rev* 1950 Jan;78(1):1-3. [doi: [10.1175/1520-0493\(1950\)078<0001:vofeit>2.0.co;2](https://doi.org/10.1175/1520-0493(1950)078<0001:vofeit>2.0.co;2)]
34. Online Breast Cancer Prognosis Tool. Fudan Breast Cancer: shinyapps.io. URL: <https://fudanbreastcancer.shinyapps.io/DynNomapp> [accessed 2021-12-15]
35. Rahman SA, Walker RC, Maynard N, Trudgill N, Crosby T, Cromwell DA, NOGCA project team AUGIS. The AUGIS Survival Predictor: Prediction of Long-term and Conditional Survival after Esophagectomy Using Random Survival Forests. *Ann Surg* 2021 Feb 17:online ahead of print. [doi: [10.1097/SLA.0000000000004794](https://doi.org/10.1097/SLA.0000000000004794)] [Medline: [33630434](https://pubmed.ncbi.nlm.nih.gov/33630434/)]

## Abbreviations

**AUC:** area under the curve

**C-index:** concordance index

**EN:** elastic net

**LASSO:** least absolute shrinkage and selection operator

**RSF:** random survival forest

**SVM:** support vector machine

*Edited by C Lovis; submitted 08.09.21; peer-reviewed by X Dong, K Rathi, JA Benítez-Andrades; comments to author 14.11.21; revised version received 15.12.21; accepted 02.01.22; published 18.02.22.*

*Please cite as:*

*Xiao J, Mo M, Wang Z, Zhou C, Shen J, Yuan J, He Y, Zheng Y*

*The Application and Comparison of Machine Learning Models for the Prediction of Breast Cancer Prognosis: Retrospective Cohort Study*

*JMIR Med Inform* 2022;10(2):e33440

URL: <https://medinform.jmir.org/2022/2/e33440>

doi: [10.2196/33440](https://doi.org/10.2196/33440)

PMID: [35179504](https://pubmed.ncbi.nlm.nih.gov/35179504/)

©Jialong Xiao, Miao Mo, Zezhou Wang, Changming Zhou, Jie Shen, Jing Yuan, Yulian He, Ying Zheng. Originally published in *JMIR Medical Informatics* (<https://medinform.jmir.org>), 18.02.2022. This is an open-access article distributed under the terms of the Creative Commons Attribution License (<https://creativecommons.org/licenses/by/4.0/>), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in *JMIR Medical Informatics*, is properly cited. The complete bibliographic information, a link to the original publication on <https://medinform.jmir.org/>, as well as this copyright and license information must be included.

Original Paper

# Automatically Explaining Machine Learning Predictions on Severe Chronic Obstructive Pulmonary Disease Exacerbations: Retrospective Cohort Study

Siyang Zeng<sup>1</sup>, MS; Mehrdad Arjomandi<sup>2,3</sup>, MD; Gang Luo<sup>1</sup>, DPhil

<sup>1</sup>Department of Biomedical Informatics and Medical Education, University of Washington, Seattle, WA, United States

<sup>2</sup>Medical Service, San Francisco Veterans Affairs Medical Center, San Francisco, CA, United States

<sup>3</sup>Department of Medicine, University of California, San Francisco, CA, 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](mailto:gangluo@cs.wisc.edu)

## Abstract

**Background:** Chronic obstructive pulmonary disease (COPD) is a major cause of death and places a heavy burden on health care. To optimize the allocation of precious preventive care management resources and improve the outcomes for high-risk patients with COPD, we recently built the most accurate model to date to predict severe COPD exacerbations, which need inpatient stays or emergency department visits, in the following 12 months. Our model is a machine learning model. As is the case with most machine learning models, our model does not explain its predictions, forming a barrier for clinical use. Previously, we designed a method to automatically provide rule-type explanations for machine learning predictions and suggest tailored interventions with no loss of model performance. This method has been tested before for asthma outcome prediction but not for COPD outcome prediction.

**Objective:** This study aims to assess the generalizability of our automatic explanation method for predicting severe COPD exacerbations.

**Methods:** The patient cohort included all patients with COPD who visited the University of Washington Medicine facilities between 2011 and 2019. In a secondary analysis of 43,576 data instances, we used our formerly developed automatic explanation method to automatically explain our model's predictions and suggest tailored interventions.

**Results:** Our method explained the predictions for 97.1% (100/103) of the patients with COPD whom our model correctly predicted to have severe COPD exacerbations in the following 12 months and the predictions for 73.6% (134/182) of the patients with COPD who had  $\geq 1$  severe COPD exacerbation in the following 12 months.

**Conclusions:** Our automatic explanation method worked well for predicting severe COPD exacerbations. After further improving our method, we hope to use it to facilitate future clinical use of our model.

**International Registered Report Identifier (IRRID):** RR2-10.2196/13783

(*JMIR Med Inform* 2022;10(2):e33043) doi:[10.2196/33043](https://doi.org/10.2196/33043)

**KEYWORDS**

chronic obstructive pulmonary disease; forecasting; machine learning; patient care management

## Introduction

### Background

Chronic obstructive pulmonary disease (COPD) is a leading cause of death [1] and affects 6.5% of American adults [2]. In the United States, COPD leads to 0.7 million inpatient stays and 1.5 million emergency department (ED) visits every year [2]. Severe COPD exacerbations are exacerbations that need inpatient stays or ED visits [3]. These exacerbations often result in irreversible deterioration in health status and lung function [4-9] and account for 90.3% of the US \$32.1 billion total annual medical costs of the United States associated with COPD [2,10]. Many of these exacerbations, which include 47% of inpatient stays and many ED visits because of COPD, are regarded as preventable with suitable outpatient care [3,11]. To reduce severe COPD exacerbations, many health care systems and health plans use predictive models to identify high-risk patients [12] for preventive care management [13]. Once a patient is enrolled in the care management program, care managers will regularly follow up with the patient on the phone to assess the patient's health status and help schedule health and related services. For patients with COPD, successful care management can cut up to 40% of their inpatient stays [14] and 27% of their ED visits [15].

As a care management program can take  $\leq 3\%$  of patients because of resource limits [16], the effectiveness of the program depends critically on the performance of the predictive model that is used. To optimize the allocation of precious care management resources and improve the outcomes for high-risk patients with COPD, we recently built the most accurate model to date to predict severe COPD exacerbations in the following 12 months [17]. Our model achieved an area under the receiver operating characteristic curve of 0.866, a sensitivity of 56.6% (103/182), and a specificity of 91.17% (6698/7347). In comparison, to the best of our knowledge, each published prior model for this prediction target [18-51] had an area under the receiver operating characteristic curve  $\leq 0.809$  and a sensitivity  $< 50\%$  when the specificity was set at approximately 91%. Our model is based on the machine learning algorithm of extreme gradient boosting (XGBoost) [52]. As is the case with most machine learning models, our model does not explain its predictions, forming a barrier for clinical use [53]. Offering explanations is essential for care managers to make sense of and trust the model's predictions to make care management enrollment decisions and identify suitable interventions. Currently, there is no consensus on what explanation means for machine learning predictions. In this paper, by explaining the prediction that a machine learning model makes on a patient, we mean to find  $\geq 1$  rule whose left-hand side is fulfilled by the patient and whose right-hand side is consistent with the prediction. Previously, we developed a method to automatically provide rule-type explanations for any machine learning model's predictions on tabular data and suggest tailored interventions with no loss of model performance [54-58]. This method has been tested before for asthma outcome prediction but not for COPD outcome prediction.

### Objective

The goal of this particular study is to assess the generalizability of our automatic explanation method for predicting severe COPD exacerbations. After further improving our method in the future, our eventual goal is that care managers can use our method to make COPD care management enrollment and intervention decisions more quickly and reliably.

## Methods

### Ethics Approval and Study Design

The institutional review board of the University of Washington Medicine (UWM) approved this retrospective cohort study (STUDY00000118) using administrative and clinical data.

### Patient Population

In Washington state, the UWM is the largest academic health care system. The enterprise data warehouse of the UWM contains administrative and clinical data from 12 clinics and 3 hospitals. This study used the same patient cohort as our previous predictive modeling study [17]. The patient cohort included all patients with COPD who visited the UWM facilities between 2011 and 2019. As adapted from the literature [59-62], a patient was deemed to have COPD if the patient was aged at least 40 years and met at least one of the following criteria:

1. The patient had "an outpatient visit diagnosis code of COPD (International Classification of Diseases, Ninth Revision (ICD-9): 491.22, 491.21, 491.9, 491.8, 493.2x, 492.8, 496; International Classification of Diseases, Tenth Revision (ICD-10): J42, J41.8, J44.\*, J43.\*) followed by  $\geq 1$  prescription of long-acting muscarinic antagonist (aclidinium, glycopyrrolate, tiotropium, and umeclidinium) within 6 months"
2. The patient had " $\geq 1$  ED or  $\geq 2$  outpatient visit diagnosis codes of COPD (International Classification of Diseases, Ninth Revision: 491.22, 491.21, 491.9, 491.8, 493.2x, 492.8, 496; International Classification of Diseases, Tenth Revision: J42, J41.8, J44.\*, J43.\*)"
3. The patient had " $\geq 1$  inpatient stay discharge having a principal diagnosis code of COPD (International Classification of Diseases, Ninth Revision: 491.22, 491.21, 491.9, 491.8, 493.2x, 492.8, 496; International Classification of Diseases, Tenth Revision: J42, J41.8, J44.\*, J43.\*)"
4. The patient had " $\geq 1$  inpatient stay discharge having a principal diagnosis code of respiratory failure (International Classification of Diseases, Ninth Revision: 518.82, 518.81, 799.1, 518.84; International Classification of Diseases, Tenth Revision: J96.0\*, J80, J96.9\*, J96.2\*, R09.2) and a secondary diagnosis code of acute COPD exacerbation (International Classification of Diseases, Ninth Revision: 491.22, 491.21, 493.22, 493.21; International Classification of Diseases, Tenth Revision: J44.1, J44.0)" [17].

We used one exclusion criterion: when calculating the data instances in a given year, the patients who died or had no encounter at the UWM during that year were excluded.

## Data Set

This study used the same structured data set as our previous predictive model paper [17]. The data set contained the administrative and clinical data of the patient cohort's encounters at the 12 UWM clinics and 3 UWM hospitals between 2011 and 2020.

## Prediction Target (Dependent or Outcome Variable)

This study used the same prediction target as our previous predictive model [17]. For a patient with COPD and  $\geq 1$  encounter at the UWM in a particular year (index year), we used patient data up to the end of the year to predict the outcome—whether the patient would have  $\geq 1$  severe COPD exacerbation in the following 12 months. A severe COPD exacerbation is defined as an inpatient stay or an ED visit with a principal diagnosis of COPD (International Classification of Diseases, Ninth Revision: 491.22, 491.21, 491.9, 491.8, 493.2x, 492.8, 496; International Classification of Diseases, Tenth Revision: J42, J41.8, J44.\*, J43.\*).

## Data Preprocessing, Predictive Model, and Features (Independent Variables)

We applied the same methods as in our previous predictive model paper [17] to perform data preprocessing. Using the upper and lower bounds provided by a clinical expert in our team, as well as the upper and lower bounds from the Guinness World Records, we pinpointed the biologically implausible values, marked them missing, and normalized each numerical feature. Our model used 229 features and the XGBoost classification algorithm [52] to make predictions. As listed in the second table in the web-based multimedia appendix of our previous paper [17], these features were calculated on the attributes in our structured data set and covered various aspects such as vital signs, diagnoses, visits, procedures, medications, laboratory tests, and patient demographics. An example feature is the number of days since the patient had the last diagnosis of acute COPD exacerbation. Each input data instance to the predictive model contained these 229 features, corresponded to a distinct patient and index year pair, and was used to predict the outcome of the patient in the following 12 months. As in our previous predictive model paper [17], the cutoff threshold for binary classification was set at the top 10% of patients with the largest predicted risk. A care management program can take  $\leq 3\%$  of patients because of resource limits [16]. After using our model to identify the top 10% of patients with the largest predicted risk and using our automatic explanation method to explain the predictions, care managers could review patient charts, consider factors such as social dimensions, and choose  $\leq 3\%$  of patients for care management enrollment. A value of 10% was chosen to strike a balance between covering a large percentage of patients who would have  $\geq 1$  severe COPD exacerbation in the following 12 months and keeping the care managers' workload manageable.

## Review of Our Automatic Explanation Method

### Overview

Previously, we developed a method to automatically provide rule-type explanations for any machine learning model's

predictions on tabular data and suggest tailored interventions with no loss of model performance [54–58]. When creating the automatic explanation function before the prediction time, our method requires  $\geq 1$  expert in the function's design team to manually provide some information, such as marking the feature–value pairs that could have a positive correlation with the bad outcome value and compiling interventions for these feature–value pair items. This can typically be performed in a few man-hours. Once this information is obtained and stored in the function's knowledge base, our method can automatically explain the machine learning model's predictions and suggest tailored interventions at the prediction time.

### Main Idea

Our automatic explanation method [54–58] uses 2 models at the same time to separate making predictions and providing explanations. Each model plays a different role. The first model is used to predict the outcome. This model can be any model that takes continuous and categorical features as its inputs and is typically chosen to be the model that performs the best at making predictions. The second model comprises class-based association rules [63,64] mined from the training set. We use the second model to explain the first model's predictions rather than to make predictions. After we convert each continuous feature into  $\geq 1$  categorical feature via automatic discretization [63,65], the association rules are mined using the Apriori algorithm, whereas other standard methods such as frequent pattern growth can also be used [64]. Every rule shows that a feature pattern links to a value  $z$  of the outcome variable in the form of:

$$p_1 \text{ AND } p_2 \text{ AND } \dots \text{ AND } p_k \rightarrow z. \quad (1)$$

Here, each item  $p_i$  ( $1 \leq i \leq k$ ) is a feature–value pair  $(x, c)$ , indicating that feature  $x$  has a value  $c$  if  $c$  is a value or a value within  $c$  if  $c$  is a range. The values of  $k$  and  $z$  can vary by rules. For the binary classification of good versus bad outcomes,  $z$  is usually the bad outcome value. The rule indicates that a patient's outcome tends to take the value  $z$  if the patient satisfies all of  $p_1, p_2, \dots$ , and  $p_k$ . The following is an example of a rule:

The patient's last diagnosis of acute COPD exacerbation was from the past 81.4 days AND the patient's COPD reliever prescriptions in the past year included  $>10$  distinct medications  $\rightarrow$  The patient will probably have at least one severe COPD exacerbation in the following 12 months.

### Mining and Pruning Rules

Each rule has two quality measures: commonality and confidence. For a rule:

$$p_1 \text{ AND } p_2 \text{ AND } \dots \text{ AND } p_k \rightarrow z, \quad (1)$$

its commonality is defined as the percentage of data instances satisfying  $p_1, p_2, \dots$ , and  $p_k$  among all the data instances linked to  $z$ . Its confidence is defined as the percentage of data instances linked to  $z$  among all the data instances satisfying  $p_1, p_2, \dots$ , and  $p_k$ . Commonality measures the coverage of a rule within the context of  $z$ . Confidence measures the precision of a rule.



The process of mining and pruning rules is controlled by five parameters: the number of top features that are used to form rules, upper limit of the number of items on the left-hand side of a rule, lower limit of confidence, lower limit of commonality, and upper limit of the confidence difference. Our method uses rules that each contains at most the upper limit number of items on its left-hand side, has a commonality that is greater than or equal to the lower limit of commonality, and has a confidence that is greater than or equal to the lower limit of confidence.

Our automatic explanation method is intended to be used for real-time clinical decision support. Once the first model provides its predicted outcome of a patient, we need to use the second model to provide automatic explanations for the prediction quickly, ideally within a subsecond. For this purpose, we need to control the number of association rules in the second model to help reduce the overhead of retrieving and ranking the relevant rules at the prediction time. We used the following three techniques to cut the number of rules:

1. Some machine learning algorithms, such as XGBoost [52], automatically calculate the importance value of each feature. When the data set included many features, we used only the top few features in the first model with the highest importance values to form rules. Usually, we set the number of top features to be used to the maximum possible number without making the association rule mining process run out of memory.
2. A rule  $r_1$  was dropped if there exists another rule  $r_2$  satisfying three conditions:  $r_1$  and  $r_2$  have the same value on their right-hand sides; the items on the left-hand side of  $r_2$  are a proper subset of the items on the left-hand side of  $r_1$  (ie,  $r_2$  is more general than  $r_1$ ); and the confidence of  $r_2$  is greater than or equal to the confidence of  $r_1$ —the upper limit of the confidence difference.
3. All distinct feature–value pairs were examined and labeled by a clinical expert in the automatic explanation function’s design team. When forming rules, we used only those feature–value pairs that the clinical expert deemed could have a positive correlation with the bad outcome value.

For every feature-value pair item used to form association rules, a clinical expert in the automatic explanation function’s design team compiled  $\geq 0$  intervention. An item is termed actionable if it is associated with  $\geq 1$  intervention. These interventions are automatically attached to the rules whose left-hand sides contain this item. A rule is termed actionable if its left-hand side contains  $\geq 1$  actionable item and, in turn, is associated with  $\geq 1$  intervention. In theory, for each combination of feature–value pair items that appears on the left-hand side of  $\geq 1$  mined rule, the clinical expert could compile additional interventions to be automatically attached to the rules whose left-hand sides contain this combination if these interventions have not already been compiled for any individual feature–value pair item in the combination. In practice, we have not needed to do this for predicting severe COPD exacerbations, whereas such a need could occur in some other clinical prediction tasks in the future.

### ***Explaining the Predictions***

For each patient predicted by the first model to have a bad outcome, we explained the prediction by presenting the association rules in the second model whose left-hand sides are fulfilled by the patient and whose right-hand sides have the bad outcome value. The rules were sorted using the method given in our paper [57]. This method incorporates 5 factors into a rule-scoring function, striking a balance among them. These factors include confidence, commonality, number of items on the left-hand side of the rule, whether the rule is actionable, and the degree of information redundancy with the higher-ranked rules. The rules are ranked based on the computed scores in an iterative fashion. Every rule offers an explanation for why the patient is predicted to have a bad outcome. For each actionable rule that is presented, the associated interventions are shown next to it. This helps the user of the automatic explanation function pinpoint suitable interventions for the patient. Typically, the rules in the second model provide common reasons for a patient to have a bad outcome. Although some patients could have bad outcomes because of rare reasons not covered by these rules, the second model usually explains most, although not all, of the bad outcomes correctly predicted by the first model.

### **Parameter Setting**

Our model [17] used 229 features to predict patient outcomes. In this study, we used the top 80 features that our model ranked with the highest importance values to form association rules. Regardless of whether all 229 features or only the top 80 features were used, our model had the same area under the receiver operating characteristic curve of 0.866.

As in our prior study on automatically explaining predictions of asthma outcomes on the UWM data [55], we set the upper limit of the number of items on the left-hand side of a rule to 5, the lower limit of commonality to 1%, and the lower limit of confidence to 50%. The last 2 values were commonly used to mine association rules [63], whereas commonality was essentially support computed on all the data instances linked to the bad outcome [54]. The first value struck a balance between the explanation power of our automatic explanation method and not making the rules too complex to understand. To set the upper limit value of the confidence difference, we plotted the number of association rules remaining from the rule pruning process versus the upper limit of the confidence difference. Our prior automatic explanation papers [54-56,58] showed that the number of remaining rules first decreased rapidly as the upper limit of the confidence difference increased and then slowly decreased after the upper limit of the confidence difference became large enough. The upper limit value of the confidence difference was set at a point where a further increase in the confidence difference had a minor impact on reducing the number of remaining rules.

### **Data Analysis**

#### ***Split of the Training and Test Sets***

We adopted the method from our previous predictive model paper [17] to split the entire data set into the training and test sets. As the outcomes were from the following year, the data

set contained 9 years of effective data (2011-2019) over the 10-year period of 2011 to 2020. To reflect how our predictive model and our automatic explanation method will be used in clinical practice in the future, we used the 2011 to 2018 data as the training set to train our model and compute the association rules used by our automatic explanation method and the 2019 data as the test set to assess the performance of our model and our automatic explanation method.

### ***Providing Examples of Automatic Explanations***

To give the reader a concrete feeling of the results produced by our automatic explanation method, we randomly selected 3 example patients from the patients who were correctly predicted by our model to have  $\geq 1$  severe COPD exacerbation in the following 12 months and for whom our automatic explanation method could offer  $\geq 1$  explanation. For each example patient, we listed the top 3 explanations given by our automatic explanation method.

### ***Performance Metrics***

We examined the performance of our automatic explanation method using the following performance metrics from our prior automatic explanation papers [54-56,58]. Regarding the explanation power of our automatic explanation method, a performance metric is the percentage of patients for whom our method could provide explanations among the patients with COPD who were correctly predicted by our model to have  $\geq 1$  severe COPD exacerbation in the following 12 months. We assessed both the average and median number of (actionable)

rules matching such a patient. A rule matches a patient if the patient satisfies all items on its left-hand side.

As shown by our prior automatic explanation papers [54-56,58], many rules matching a patient often differ from each other by only 1 item on their left-hand sides. In this case, the number of rules greatly exceeded the amount of nonrepeated information contained in these rules. To provide a comprehensive overview of the amount of information provided by the automatic explanations, we examined the distributions of (1) the number of (actionable) rules and (2) the number of unique actionable items in the rules matching a patient who was correctly predicted by our model to have  $\geq 1$  severe COPD exacerbation in the following 12 months.

## ***Results***

### **Characteristics of Our Patient Cohort**

Each data instance corresponds to a distinct patient and index year pair. Tables 1 and 2 summarize the patient demographic and clinical characteristics of the data instances in the training and test sets, respectively. These 2 sets of characteristics were relatively similar to each other. In the training set, 5.66% (2040/36,047) of the data instances were related to severe COPD exacerbations in the following 12 months. In the test set, 2.42% (182/7529) of the data instances were related to severe COPD exacerbations in the following 12 months. A detailed comparison of these 2 sets of characteristics was provided in our previous predictive model paper [17].

**Table 1.** The patient demographic and clinical characteristics of the data instances in the training set.

Patient characteristics	Data instances related to no severe COPD <sup>a</sup> exacerbation in the following 12 months (n=34,007), n (%)	Data instances related to severe COPD exacerbations in the following 12 months (n=2040), n (%)	Data instances (n=36,047), n (%)
<b>Sex</b>			
Female	14,665 (43.12)	749 (36.72)	15,414 (42.76)
Male	19,342 (56.88)	1291 (63.28)	20,633 (57.24)
<b>Age (years)</b>			
40-65	17,574 (51.68)	1219 (59.75)	18,793 (52.13)
>65	16,433 (48.32)	821 (40.25)	17,254 (47.87)
<b>Race</b>			
White	26,117 (76.8)	1330 (65.2)	27,447 (76.14)
Black or African American	4271 (12.56)	524 (25.69)	4795 (13.3)
Asian	1948 (5.73)	144 (7.06)	2092 (5.8)
American Indian or Alaska Native	687 (2.02)	26 (1.27)	713 (1.98)
Native Hawaiian or other Pacific Islander	176 (0.52)	8 (0.39)	184 (0.51)
Other, unknown, or not reported	808 (2.37)	8 (0.39)	816 (2.27)
<b>Ethnicity</b>			
Hispanic	804 (2.36)	53 (2.6)	857 (2.38)
Non-Hispanic	30,644 (90.11)	1941 (95.15)	32,585 (90.39)
Unknown or not reported	2559 (7.53)	46 (2.25)	2605 (7.23)
<b>Insurance</b>			
Public	27,831 (81.84)	1767 (86.62)	29,598 (82.11)
Private	16,679 (49.05)	834 (40.88)	17,513 (48.58)
Self-paid or charity	1765 (5.19)	229 (11.23)	1994 (5.53)
<b>Number of years since the first encounter related to COPD in the data set</b>			
≤3	28,749 (84.54)	1566 (76.76)	30,315 (84.1)
>3	5258 (15.46)	474 (23.24)	5732 (15.90)
<b>Smoking status</b>			
Current smoker	15,863 (46.65)	1089 (53.38)	16,952 (47.03)
Former smoker	7022 (20.65)	345 (16.91)	7367 (20.44)
Never smoker or unknown	11,122 (32.7)	606 (29.71)	11,728 (32.53)
<b>COPD medication prescription</b>			
SABA <sup>b</sup>	20,865 (61.36)	1684 (82.55)	22,549 (62.55)
SAMA <sup>c</sup>	8566 (25.19)	1042 (51.08)	9608 (26.65)
SABA and SAMA combination	6364 (18.71)	810 (39.71)	7174 (19.9)
LABA <sup>d</sup>	8062 (23.71)	842 (41.27)	8904 (24.7)
LAMA <sup>e</sup>	9242 (27.18)	1001 (49.07)	10,243 (28.42)
LABA and LAMA combination	386 (1.14)	40 (1.96)	426 (1.18)
ICS <sup>f</sup>	12,208 (35.9)	1119 (54.85)	13,327 (36.97)
ICS and LABA combination	7544 (22.18)	782 (38.33)	8326 (23.1)
ICS, LABA, and LAMA combination	16 (0.05)	0 (0)	16 (0.04)

Patient characteristics	Data instances related to no severe COPD <sup>a</sup> exacerbation in the following 12 months (n=34,007), n (%)	Data instances related to severe COPD exacerbations in the following 12 months (n=2040), n (%)	Data instances (n=36,047), n (%)
Systemic corticosteroid	10,149 (29.84)	1144 (56.08)	11,293 (31.33)
Phosphodiesterase-4 inhibitor	84 (0.25)	10 (0.49)	94 (0.26)
<b>Comorbidity</b>			
Anxiety or depression	10,061 (29.59)	725 (35.54)	10,786 (29.92)
Allergic rhinitis	2271 (6.68)	174 (8.53)	2445 (6.78)
Asthma	4377 (12.87)	417 (20.44)	4794 (13.3)
Diabetes	7177 (21.1)	446 (21.86)	7623 (21.15)
Congestive heart failure	5568 (16.37)	495 (24.26)	6063 (16.82)
Eczema	1460 (4.29)	98 (4.8)	1558 (4.32)
Hypertension	17,211 (50.61)	1150 (56.37)	18,361 (50.94)
Gastroesophageal reflux	6655 (19.57)	507 (24.85)	7162 (19.87)
Ischemic heart disease	6934 (20.39)	486 (23.82)	7420 (20.58)
Obesity	3232 (9.5)	255 (12.5)	3487 (9.67)
Lung cancer	742 (2.18)	52 (2.55)	794 (2.2)
Sleep apnea	2926 (8.6)	253 (12.4)	3179 (8.82)
Sinusitis	1299 (3.82)	83 (4.07)	1382 (3.83)

<sup>a</sup>COPD: chronic obstructive pulmonary disease.

<sup>b</sup>SABA: short-acting beta-2 agonist.

<sup>c</sup>SAMA: short-acting muscarinic antagonist.

<sup>d</sup>LABA: long-acting beta-2 agonist.

<sup>e</sup>LAMA: long-acting muscarinic antagonist.

<sup>f</sup>ICS: inhaled corticosteroid.

**Table 2.** The patient demographic and clinical characteristics of the data instances in the test set.

Patient characteristics	Data instances related to no severe COPD <sup>a</sup> exacerbation in the following 12 months (n=7347), n (%)	Data instances related to severe COPD exacerbations in the following 12 months (n=182), n (%)	Data instances (n=7529), n (%)
<b>Sex</b>			
Female	3242 (44.13)	47 (25.8)	3289 (43.68)
Male	4105 (55.87)	135 (74.2)	4240 (56.32)
<b>Age (years)</b>			
40-65	3324 (45.24)	118 (64.8)	3442 (45.72)
>65	4023 (54.76)	64 (35.2)	4087 (54.28)
<b>Race</b>			
White	5682 (77.34)	111 (61.0)	5793 (76.94)
Black or African American	839 (11.42)	57 (31.3)	896 (11.9)
Asian	432 (5.88)	7 (3.9)	439 (5.83)
American Indian or Alaska Native	151 (2.06)	5 (2.7)	156 (2.07)
Native Hawaiian or other Pacific Islander	51 (0.69)	2 (1.1)	53 (0.71)
Other, unknown, or not reported	192 (2.61)	0 (0.0)	192 (2.55)
<b>Ethnicity</b>			
Hispanic	185 (2.52)	3 (1.6)	188 (2.5)
Non-Hispanic	6909 (94.04)	179 (98.4)	7088 (94.14)
Unknown or not reported	253 (3.44)	0 (0)	253 (3.36)
<b>Insurance</b>			
Public	6722 (91.49)	179 (98.4)	6901 (91.66)
Private	4532 (61.69)	110 (60.4)	4642 (61.65)
Self-paid or charity	499 (6.79)	41 (22.5)	540 (7.17)
<b>Number of years since the first encounter related to COPD in the data set</b>			
≤3	5073 (69.05)	81 (44.5)	5154 (68.46)
>3	2274 (30.95)	101 (55.5)	2375 (31.54)
<b>Smoking status</b>			
Current smoker	3781 (51.46)	112 (61.5)	3893 (51.71)
Former smoker	1242 (16.91)	25 (13.7)	1267 (16.83)
Never smoker or unknown	2324 (31.63)	45 (24.7)	2369 (31.47)
<b>COPD medication prescription</b>			
SABA <sup>b</sup>	4083 (55.57)	158 (86.8)	4241 (56.33)
SAMA <sup>c</sup>	1134 (15.43)	68 (37.4)	1202 (15.96)
SABA and SAMA combination	1694 (23.06)	115 (63.2)	1809 (24.03)
LABA <sup>d</sup>	1683 (22.91)	77 (42.3)	1760 (23.38)
LAMA <sup>e</sup>	1951 (26.56)	110 (60.4)	2061 (27.37)
LABA and LAMA combination	388 (5.28)	12 (6.6)	400 (5.31)
ICS <sup>f</sup>	2537 (34.53)	98 (53.8)	2635 (35)
ICS and LABA combination	1729 (23.53)	75 (41.2)	1804 (23.96)
ICS, LABA, and LAMA combination	68 (0.93)	1 (0.5)	69 (0.92)

Patient characteristics	Data instances related to no severe COPD <sup>a</sup> exacerbation in the following 12 months (n=7347), n (%)	Data instances related to severe COPD exacerbations in the following 12 months (n=182), n (%)	Data instances (n=7529), n (%)
Systemic corticosteroid	2282 (31.06)	103 (56.6)	2385 (31.68)
Phosphodiesterase-4 inhibitor	24 (0.33)	2 (1.1)	26 (0.35)
<b>Comorbidity</b>			
Anxiety or depression	2090 (28.45)	63 (34.6)	2153 (28.6)
Allergic rhinitis	396 (5.39)	14 (7.7)	410 (5.45)
Asthma	1053 (14.33)	43 (23.6)	1096 (14.56)
Diabetes	1649 (22.44)	40 (22)	1689 (22.43)
Congestive heart failure	1369 (18.63)	43 (23.6)	1412 (18.75)
Eczema	247 (3.36)	11 (6)	258 (3.43)
Hypertension	3686 (50.17)	105 (57.7)	3791 (50.35)
Gastroesophageal reflux	1396 (19)	47 (25.8)	1443 (19.17)
Ischemic heart disease	1604 (21.83)	54 (29.7)	1658 (22.02)
Obesity	648 (8.82)	21 (11.5)	669 (8.89)
Lung cancer	200 (2.72)	3 (1.6)	203 (2.7)
Sleep apnea	887 (12.07)	28 (15.4)	915 (12.15)
Sinusitis	272 (3.7)	7 (3.8)	279 (3.71)

<sup>a</sup>COPD: chronic obstructive pulmonary disease.

<sup>b</sup>SABA: short-acting beta-2 agonist.

<sup>c</sup>SAMA: short-acting muscarinic antagonist.

<sup>d</sup>LABA: long-acting beta-2 agonist.

<sup>e</sup>LAMA: long-acting muscarinic antagonist.

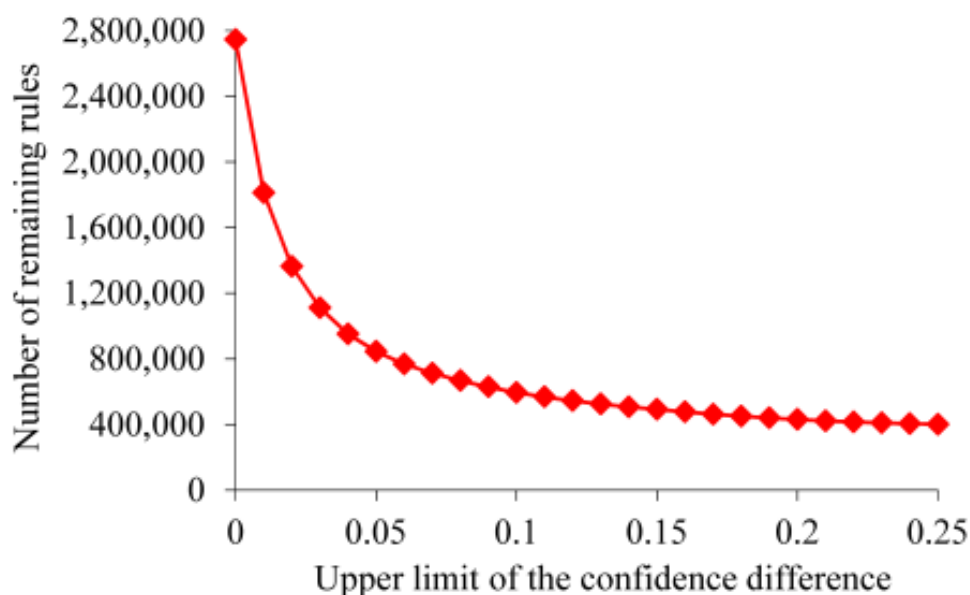
<sup>f</sup>ICS: inhaled corticosteroid.

### The Number of Association Rules

Using the top 80 features ranked with the highest importance values in our predictive model, 7,729,134 association rules were mined from the training set. Figure 1 shows the number of remaining rules versus the upper limit of the confidence difference

difference. The number of remaining rules first rapidly decreased as the upper limit of the confidence difference increased and then slowly decreased after the upper limit of the confidence difference became  $\geq 0.15$ . We set the upper limit of the confidence difference to the value of 0.15, resulting in 492,803 remaining rules.

**Figure 1.** The number of remaining association rules versus the upper limit of the confidence difference.



The top 80 features totally had 219 distinct feature–value pairs, 141 (64.4%) of which were actionable. A clinical expert on COPD (MA) in our team reviewed all distinct feature–value pairs of the top 80 features and labeled those that could have a positive correlation with severe COPD exacerbations in the following 12 months. After dropping the rules containing any other feature–value pair items, 460,592 rules were left. These rules were all actionable.

### Examples of the Produced Automatic Explanations

To give the reader a concrete feeling of the results produced by our automatic explanation method, we randomly selected 3 example patients from the patients who were correctly predicted by our model to have  $\geq 1$  severe COPD exacerbation in the following 12 months and for whom our automatic explanation method could offer  $\geq 1$  explanation. [Tables 3-5](#) show the top 3 explanations that our automatic explanation method provided for every example patient.

**Table 3.** The top 3 association rules generated for the first example patient.

Rank, rule, and item on the rule's left-hand side	Interpretation of the item	Interventions linked to the item
<b>Rank 1: The patient's last diagnosis of acute COPD<sup>a</sup> exacerbation was from the past 81.4 days AND the patient's COPD reliever prescriptions in the past year included &gt;10 distinct medications → the patient will probably have at least one severe COPD exacerbation in the following 12 months</b>		
The patient's last diagnosis of acute COPD exacerbation was from the past 81.4 days	Having a recent acute COPD exacerbation shows a need for better control of the disease.	<ul style="list-style-type: none"> <li>• Provide education on managing COPD and more frequent follow-ups</li> <li>• Ensure use of appropriate COPD medications</li> <li>• Consider influenza shot, pneumonia vaccination, or smoking cessation</li> <li>• Assess the need for pulmonary rehabilitation or home care</li> <li>• Ensure that the patient has a primary care provider or is referred to a specialist</li> </ul>
The patient's COPD reliever prescriptions in the past year included >10 distinct medications	Using many rescue medications for COPD indicates ineffective regimen, poor treatment adherence, or poor control of the disease.	<ul style="list-style-type: none"> <li>• Simplify COPD medications to once-a-day formulations or combination medications</li> <li>• Address concerns for adverse interactions between medications</li> <li>• Provide education on the correct use of COPD medications or inhalers</li> <li>• Consider strategies to improve medication adherence such as providing reminders for taking medications in time</li> <li>• Medication reconciliation review by a physician or a pharmacist</li> </ul>
<b>Rank 2: The patient had between 8 and 19 diagnoses of acute COPD exacerbation in the past year AND the patient's last COPD diagnosis was from the past 25.6 days AND the patient's nebulizer medication prescriptions in the past year included &gt;11 medications → the patient will probably have at least one severe COPD exacerbation in the following 12 months</b>		
The patient had between 8 and 19 diagnoses of acute COPD exacerbation in the past year	Frequently having acute COPD exacerbations shows a need for better control of the disease.	<ul style="list-style-type: none"> <li>• Provide education on managing COPD and more frequent follow-ups</li> <li>• Ensure use of appropriate COPD medications</li> <li>• Consider influenza shot, pneumonia vaccination, or smoking cessation</li> <li>• Assess the need for pulmonary rehabilitation or home care</li> </ul>
The patient's last COPD diagnosis was from the past 25.6 days	Having a recent COPD diagnosis associated with an ED <sup>b</sup> visit or an inpatient stay indicates poor control of the disease.	<ul style="list-style-type: none"> <li>• Provide education on managing COPD and more frequent follow-ups</li> <li>• Ensure use of appropriate COPD medications</li> <li>• Consider influenza shot, pneumonia vaccination, or smoking cessation</li> <li>• Assess the need for pulmonary rehabilitation or home care</li> </ul>
The patient's nebulizer medication prescriptions in the past year included >11 medications	Using many medications for COPD with a nebulizer indicates an ineffective regimen, poor treatment adherence, or poor control of the disease. Using nebulizer medications could be a sign of having a mild exacerbation or more severe COPD.	<ul style="list-style-type: none"> <li>• Simplify COPD medications to once-a-day formulations or combination medications</li> <li>• Address concerns for adverse interactions between medications</li> <li>• Provide education on the correct use of COPD medications or inhalers</li> <li>• Consider strategies to improve medication adherence such as providing reminders for taking medications in time</li> <li>• Medication reconciliation review by a physician or a pharmacist</li> </ul>
<b>Rank 3: The patient's average length of an inpatient stay in the past year was between 0.61 and 7.66 days AND the patient's last outpatient visit on COPD occurred in the past 82.4 days AND the patient's nebulizer medication prescriptions in the past year included &gt;11 medications AND the patient's maximum percentage of neutrophils in the past year was &gt;76.5% → the patient will probably have at least one severe COPD exacerbation in the following 12 months</b>		
The patient's average length of an inpatient stay in the past year was between 0.61 and 7.66 days	Having a long inpatient stay can indicate that the patient has a more severe disease or comorbidities.	<ul style="list-style-type: none"> <li>• Ensure that the patient has a primary care provider</li> <li>• Assess the need for home care or referral to a skilled nursing facility</li> <li>• Provide education on managing COPD and resources for care</li> <li>• Ensure use of appropriate COPD medications</li> </ul>
The patient's last outpatient visit on COPD occurred in the past 82.4 days	If the patient's last outpatient visit on COPD was for acute problems with COPD, it could indicate poor control of the disease and a need for additional support to control COPD.	<ul style="list-style-type: none"> <li>• Provide education on managing COPD and resources for care</li> <li>• Ensure use of appropriate COPD medications</li> <li>• Assess the need for home care or pulmonary rehabilitation</li> </ul>
The patient's nebulizer medication prescriptions in the past year included >11 medications	Using many medications for COPD with a nebulizer indicates an ineffective regimen, poor treatment adherence, or poor control of the disease. Using nebulizer medications could be a sign of having a mild exacerbation or more severe COPD.	<ul style="list-style-type: none"> <li>• Simplify COPD medications to once-a-day formulations or combination medications</li> <li>• Address concerns for adverse interactions between medications</li> <li>• Provide education on the correct use of COPD medications or inhalers</li> <li>• Consider strategies to improve medication adherence such as providing reminders for taking medications in time</li> <li>• Medication reconciliation review by a physician or a pharmacist</li> </ul>



Rank, rule, and item on the rule's left-hand side	Interpretation of the item	Interventions linked to the item
The patient's maximum percentage of neutrophils in the past year was >76.5%	Having a large percentage of neutrophils can indicate infections or distress.	<ul style="list-style-type: none"><li>• Evaluate the respiratory system, for example, using radiographic imaging</li><li>• Consider doing diagnostic tests such as viral panel, sputum culture, or procalcitonin</li><li>• Evaluate other potential morbidities such as cardiovascular disease with an electrocardiogram, echocardiography, or laboratory tests such as brain natriuretic peptide or D-dimer</li></ul>

<sup>a</sup>COPD: chronic obstructive pulmonary disease.

<sup>b</sup>ED: emergency department.

**Table 4.** The top 3 association rules generated for the second example patient.

Rank, rule, and item on the rule's left-hand side	Interpretation of the item	Interventions linked to the item
<b>Rank 1: The patient's last diagnosis of acute COPD<sup>a</sup> exacerbation was from the past 81.4 days AND the patient had &gt;2 ED<sup>b</sup> visits in the past 6 months AND the patient's nebulizer medication prescriptions in the past year included &gt;11 medications → the patient will probably have at least one severe COPD exacerbation in the following 12 months</b>		
The patient's last diagnosis of acute COPD exacerbation was from the past 81.4 days	Having a recent acute COPD exacerbation shows a need for better control of the disease.	<ul style="list-style-type: none"> <li>• Provide education on managing COPD and more frequent follow-ups</li> <li>• Ensure use of appropriate COPD medications</li> <li>• Consider influenza shot, pneumonia vaccination, or smoking cessation</li> <li>• Assess the need for pulmonary rehabilitation or home care</li> <li>• Ensure that the patient has a primary care provider or is referred to a specialist</li> </ul>
The patient had >2 ED visits in the past 6 months	Using the ED indicates poor control of conditions or a lack of access to primary, specialty, or home care.	<ul style="list-style-type: none"> <li>• Provide education on managing COPD and more frequent follow-ups</li> <li>• Ensure use of appropriate COPD medications</li> <li>• Consider influenza shot, pneumonia vaccination, or smoking cessation</li> <li>• Assess the need for pulmonary rehabilitation or home care</li> <li>• Ensure that the patient has a primary care provider or is referred to a specialist</li> </ul>
The patient's nebulizer medication prescriptions in the past year included >11 medications	Using many medications for COPD with a nebulizer indicates an ineffective regimen, poor treatment adherence, or poor control of the disease. Using nebulizer medications could be a sign of having a mild exacerbation or more severe COPD.	<ul style="list-style-type: none"> <li>• Simplify COPD medications to once-a-day formulations or combination medications</li> <li>• Address concerns for adverse interactions between medications</li> <li>• Provide education on the correct use of COPD medications or inhalers</li> <li>• Consider strategies to improve medication adherence such as providing reminders for taking medications in time</li> <li>• Medication reconciliation review by a physician or a pharmacist</li> </ul>
<b>Rank 2: The patient's maximum BMI in the past year was &lt;22.81 AND the patient's last ED visit related to COPD occurred no less than 27.2 days ago and no more than 94.3 days ago AND the patient's average length of stay of an ED visit in the past year was between 0.03 and 0.29 day AND the patient had between 2 and 4 encounters related to acute COPD exacerbation or respiratory failure in the past year → the patient will probably have at least one severe COPD exacerbation in the following 12 months</b>		
The patient's maximum BMI in the past year was <22.81	Having an unintentional weight loss can indicate comorbidities or other complications, such as malnutrition or metabolic syndrome.	<ul style="list-style-type: none"> <li>• Optimize nutritional status to address low BMI</li> <li>• Provide dietary education and advise appropriate exercise</li> </ul>
The patient's last ED visit related to COPD occurred no less than 27.2 days ago and no more than 94.3 days ago	Having a recent ED visit related to COPD shows a need for better control of the disease.	<ul style="list-style-type: none"> <li>• Provide education on managing COPD and more frequent follow-ups</li> <li>• Ensure use of appropriate COPD medications</li> <li>• Consider influenza shot, pneumonia vaccination, or smoking cessation</li> <li>• Assess the need for pulmonary rehabilitation or home care</li> <li>• Ensure that the patient has a primary care provider or is referred to a specialist</li> </ul>
The patient's average length of stay of an ED visit in the past year was between 0.03 and 0.29 day	Using the ED indicates poor control of conditions or a lack of access to primary, specialty, or home care.	<ul style="list-style-type: none"> <li>• Provide education on managing COPD and more frequent follow-ups</li> <li>• Ensure use of appropriate COPD medications</li> <li>• Consider influenza shot, pneumonia vaccination, or smoking cessation</li> <li>• Assess the need for pulmonary rehabilitation or home care</li> <li>• Ensure that the patient has a primary care provider or is referred to a specialist</li> </ul>
The patient had between 2 and 4 encounters related to acute COPD exacerbation or respiratory failure in the past year	Frequently having acute COPD exacerbations or respiratory failures shows a need for better control of the disease.	<ul style="list-style-type: none"> <li>• Provide education on managing COPD and more frequent follow-ups</li> <li>• Ensure use of appropriate COPD medications</li> <li>• Consider influenza shot, pneumonia vaccination, or smoking cessation</li> <li>• Assess the need for pulmonary rehabilitation or home care</li> <li>• Ensure that the patient has a primary care provider or is referred to a specialist</li> </ul>
<b>Rank 3: The patient had between 3 and 5 ED visits in the past year AND the patient's minimum SpO<sub>2</sub><sup>c</sup> in the past year was between 17% and 89.5% AND the patient's maximum percentage of neutrophils in the past year was &gt;76.5% AND the patient smoked &gt;0.48 pack of cigarettes per day in the past year → the patient will probably have at least one severe COPD exacerbation in the following 12 months</b>		

Rank, rule, and item on the rule's left-hand side	Interpretation of the item	Interventions linked to the item
The patient had between 3 and 5 ED visits in the past year	Using the ED indicates poor control of conditions or a lack of access to primary, specialty, or home care.	<ul style="list-style-type: none"> <li>• Provide education on managing COPD and more frequent follow-ups</li> <li>• Ensure use of appropriate COPD medications</li> <li>• Consider influenza shot, pneumonia vaccination, or smoking cessation</li> <li>• Assess the need for pulmonary rehabilitation or home care</li> <li>• Ensure that the patient has a primary care provider or is referred to a specialist</li> </ul>
The patient's minimum SpO <sub>2</sub> in the past year was between 17% and 89.5%	Having a low SpO <sub>2</sub> indicates worsening of symptoms or other complications such as hypoxemia.	<ul style="list-style-type: none"> <li>• Evaluate for cardiopulmonary causes of hypoxemia</li> <li>• Consider nighttime oximetry or sleep study to evaluate for nighttime hypoxemia or sleep apnea</li> <li>• Assess the need for home oxygen or nighttime noninvasive ventilation</li> </ul>
The patient's maximum percentage of neutrophils in the past year was >76.5%	Having a large percentage of neutrophils can indicate infections or distress.	<ul style="list-style-type: none"> <li>• Evaluate the respiratory system, for example, using radiographic imaging</li> <li>• Consider doing diagnostic tests such as viral panel, sputum culture, or procalcitonin</li> <li>• Evaluate other potential morbidities such as cardiovascular disease with an electrocardiogram, echocardiography, or laboratory tests such as brain natriuretic peptide or D-dimer</li> </ul>
The patient smoked >0.48 pack of cigarettes per day in the past year	Smoking is a key risk factor for COPD complications.	<ul style="list-style-type: none"> <li>• Provide education on the health risks of smoking</li> <li>• Suggest and provide support for smoking cessation</li> </ul>

<sup>a</sup>COPD: chronic obstructive pulmonary disease.

<sup>b</sup>ED: emergency department.

<sup>c</sup>SpO<sub>2</sub>: peripheral capillary oxygen saturation.

**Table 5.** The top 3 association rules generated for the third example patient.

Rank, rule, and item on the rule's left-hand side	Interpretation of the item	Interventions linked to the item
<b>Rank 1: The patient had between 24 and 49 COPD<sup>a</sup> diagnoses in the past year AND the patient had &gt;11 nebulizer medication prescriptions in the past year AND the patient is Black or an African American → the patient will probably have at least one severe COPD exacerbation in the following 12 months</b>		
The patient had between 24 and 49 COPD diagnoses in the past year	Frequently receiving COPD diagnoses indicates poor control of the disease.	<ul style="list-style-type: none"> <li>• Provide education on managing COPD and more frequent follow-ups</li> <li>• Ensure use of appropriate COPD medications</li> <li>• Consider influenza shot, pneumonia vaccination, or smoking cessation</li> <li>• Assess the need for pulmonary rehabilitation or home care</li> </ul>
The patient had >11 nebulizer medication prescriptions in the past year	Using many medications for COPD with a nebulizer indicates an ineffective regimen, poor treatment adherence, or poor control of the disease. Using nebulizer medications could be a sign of having a mild exacerbation or more severe COPD.	<ul style="list-style-type: none"> <li>• Simplify COPD medications to once-a-day formulations or combination medications</li> <li>• Address concerns for adverse interactions between medications</li> <li>• Provide education on the correct use of COPD medications or inhalers</li> <li>• Consider strategies to improve medication adherence such as providing reminders for taking medications in time</li> <li>• Medication reconciliation review by a physician or a pharmacist</li> </ul>
The patient is a Black or an African American	Poor respiratory outcomes and low quality of life are more prevalent in Black and African American patients.	<ul style="list-style-type: none"> <li>• Ensure that the patient has needed resources and access to care</li> <li>• Assess the need for social work or home care</li> </ul>
<b>Rank 2: The patient's last ED<sup>b</sup> visit related to COPD occurred no less than 27.2 days ago and no more than 94.3 days ago AND the patient's COPD medication prescriptions in the past year included between 13 and 16 distinct medications AND the patient's last outpatient visit on COPD occurred no less than 82.4 days ago and no more than 327.6 days ago AND the patient's maximum percentage of neutrophils in the past year was &gt;76.5% → the patient will probably have at least one severe COPD exacerbation in the following 12 months</b>		
The patient's last ED visit related to COPD occurred no less than 27.2 days ago and no more than 94.3 days ago	Having a recent ED visit related to COPD shows a need for better control of the disease.	<ul style="list-style-type: none"> <li>• Provide education on managing COPD and more frequent follow-ups</li> <li>• Ensure use of appropriate COPD medications</li> <li>• Consider influenza shot, pneumonia vaccination, or smoking cessation</li> <li>• Assess the need for pulmonary rehabilitation or home care</li> <li>• Ensure that the patient has a primary care provider or is referred to a specialist</li> </ul>
The patient's COPD medication prescriptions in the past year included between 13 and 16 distinct medications	Using many COPD medications can indicate an ineffective regimen, poor treatment adherence, or poor control of the disease.	<ul style="list-style-type: none"> <li>• Simplify COPD medications to once-a-day formulations or combination medications</li> <li>• Address concerns for adverse interactions between medications</li> <li>• Provide education on the correct use of COPD medications or inhalers</li> <li>• Consider strategies to improve medication adherence such as using a pill organizer or providing reminders for taking medications in time</li> <li>• Medication reconciliation review by a physician or a pharmacist</li> </ul>
The patient's last outpatient visit on COPD occurred no less than 82.4 days ago and no more than 327.6 days ago	If the patient's last outpatient visit on COPD was for acute problems with COPD, it could indicate poor control of the disease and a need for additional support to control COPD.	<ul style="list-style-type: none"> <li>• Provide education on managing COPD and resources for care</li> <li>• Ensure use of appropriate COPD medications</li> <li>• Assess the need for home care</li> </ul>
The patient's maximum percentage of neutrophils in the past year was >76.5%	Having a large percentage of neutrophils can indicate infections or distress.	<ul style="list-style-type: none"> <li>• Evaluate the respiratory system, for example, using radiographic imaging</li> <li>• Consider doing diagnostic tests such as viral panel, sputum culture, or procalcitonin</li> <li>• Evaluate other potential morbidities such as cardiovascular disease with an electrocardiogram, echocardiography, or laboratory tests such as brain natriuretic peptide or D-dimer</li> </ul>
<b>Rank 3: The patient had between 8 and 19 diagnoses of acute COPD exacerbation in the past year AND the relative decline of the patient's BMI in the past year was &gt;0.44% AND the patient's total length of inpatient stays in the past year was &gt;0.6 day → the patient will probably have at least one severe COPD exacerbation in the following 12 months</b>		
The patient had between 8 and 19 diagnoses of acute COPD exacerbation in the past year	Frequently having acute COPD exacerbations shows a need for better control of the disease.	<ul style="list-style-type: none"> <li>• Provide education on managing COPD and more frequent follow-ups</li> <li>• Ensure use of appropriate COPD medications</li> <li>• Consider influenza shot, pneumonia vaccination, or smoking cessation</li> <li>• Assess the need for pulmonary rehabilitation or home care</li> <li>• Ensure that the patient has a primary care provider or is referred to a specialist</li> </ul>

Rank, rule, and item on the rule's left-hand side	Interpretation of the item	Interventions linked to the item
The relative decline of the patient's BMI in the past year was >0.44%	Having an unintentional weight loss can indicate comorbidities or other complications, such as malnutrition or metabolic syndrome.	<ul style="list-style-type: none"> <li>Optimize nutritional status to address low BMI</li> <li>Provide dietary education and advise appropriate exercise</li> </ul>
The patient's total length of inpatient stays in the past year was >0.6 day	Having a long inpatient stay can indicate that the patient has a more severe disease or comorbidities. Having frequent inpatient stays shows a need for better control of the disease.	<ul style="list-style-type: none"> <li>Ensure that the patient has a primary care provider</li> <li>Assess the need for home care or referral to a skilled nursing facility</li> <li>Provide education on managing COPD and resources for care</li> <li>Ensure use of appropriate COPD medications</li> </ul>

<sup>a</sup>COPD: chronic obstructive pulmonary disease.

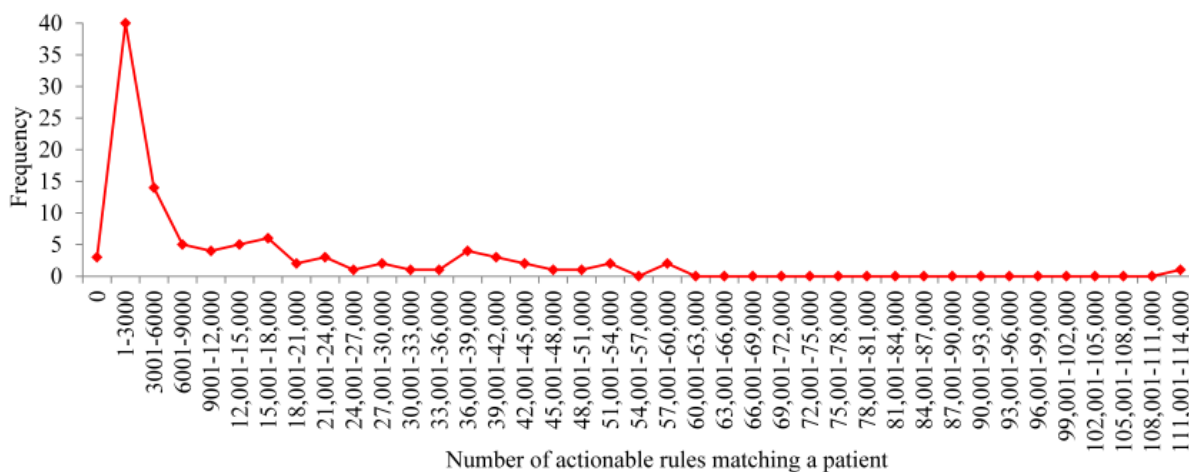
<sup>b</sup>ED: emergency department.

### Performance of the Automatic Explanation Method

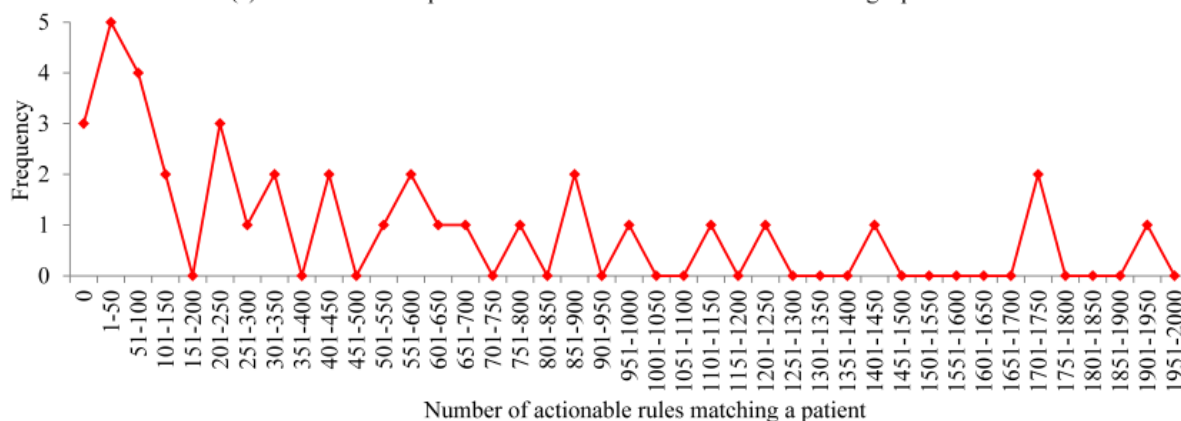
The automatic explanation method was evaluated using the test set. Our method explained the predictions for 97.1% (100/103) of the patients with COPD who were correctly predicted by our model to have severe COPD exacerbations in the following 12 months. For each such patient, our method gave an average of 13,880.19 (SD 18,700.60) explanations covering 39.80 (SD 11.98) distinct actionable items, a median of 4474 explanations, and a median of 41 distinct actionable items covered by the explanations. Each explanation corresponds to an association rule.

For the patients with COPD who were correctly predicted by our model to have severe COPD exacerbations in the following 12 months, Figure 2 shows the distribution of the number of actionable rules matching a patient. This distribution is highly skewed toward the left with a long tail. As the number of actionable rules matching a patient increases, the frequency of cases in the corresponding equal-width bucket tends to rapidly decrease in a nonmonotonic way. The largest number of actionable rules matching a patient is rather large (111,062). Nevertheless, only 1 patient matches so many rules.

**Figure 2.** The distribution of the number of actionable rules matching a patient who was correctly predicted by our model to have ≥1 severe chronic obstructive pulmonary disease exacerbation in the following 12 months.



(a) When no limit is put on the number of actionable rules matching a patient.

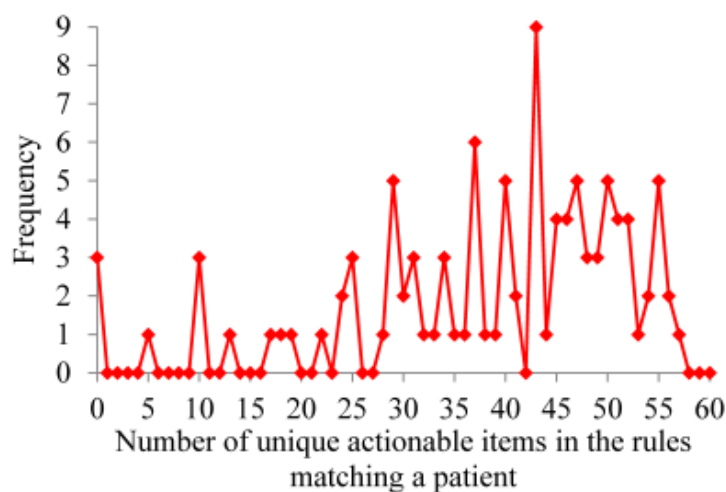


(b) When the number of actionable rules matching a patient is ≤2000.

For the patients with COPD who were correctly predicted by our model to have severe COPD exacerbations in the following 12 months, Figure 3 shows the distribution of the number of unique actionable items in the rules matching a patient. The largest number of unique actionable items in the rules matching

a patient is 57, which is much smaller than the largest number of actionable rules matching a patient. As shown in Tables 3-5, the same intervention could be linked to  $\geq 1$  distinct actionable item in the rules matching a patient.

**Figure 3.** The distribution of the number of unique actionable items in the rules matching a patient who was correctly predicted by our model to have  $\geq 1$  severe chronic obstructive pulmonary disease exacerbation in the following 12 months.



Our automatic explanation method explained the predictions for 73.6% (134/182) of the patients with COPD who had  $\geq 1$  severe COPD exacerbation in the following 12 months.

## Discussion

### Principal Findings

Our automatic explanation method generalizes well in predicting severe COPD exacerbations. Our method explained the predictions for 97.1% (100/103) of the patients with COPD who were correctly predicted by our model to have severe COPD exacerbations in the following 12 months. This percentage is comparable with the corresponding percentages of 87.6% to 97.6% that we previously obtained to explain the predictions of asthma outcomes [54-56]. This percentage is sufficiently large to apply our automatic explanation method to routine clinical use for COPD management. After further improving the performance of our model for predicting severe COPD exacerbations and our automatic explanation method, we hope our model can be used in conjunction with our automatic explanation method to provide decision support for allocating COPD care management resources and improve outcomes.

Our automatic explanation method explained the predictions for 73.6% (134/182) of the patients with COPD who had  $\geq 1$  severe COPD exacerbation in the following 12 months. This percentage is <97.1% (100/103), the success rate at which our method explained the predictions for the patients with COPD whom our model correctly predicted to have severe COPD exacerbations in the following 12 months. This seems likely to be because of the correlation between the prediction results of our model and the association rules. Among the patients whom our model correctly predicted to have severe COPD exacerbations in the following 12 months, many seem to be easy cases for using association rules to explain the outcomes. Among the patients who had severe COPD exacerbations but

were incorrectly predicted by our model to have no severe COPD exacerbation in the following 12 months, many seem to be difficult cases for any model to correctly predict or explain the outcomes.

### Related Work

Several years ago, we designed our automatic explanation method to handle relatively balanced data and demonstrated our method for predicting the diagnosis of type 2 diabetes [58]. Later, other researchers demonstrated our method on several other clinical predictive modeling tasks, such as predicting lung transplantation or mortality in patients with cystic fibrosis [66] and predicting cardiac mortality in patients with cancer [67]. Recently, we extended our automatic explanation method so it can also handle imbalanced data, where one value of the outcome variable appears much less often than another. We demonstrated our extended method for predicting hospital encounters for asthma in patients with asthma in 3 health care systems separately [54-56]. Imbalanced data also appear in the case of predicting severe COPD exacerbations, which is the use case of this paper.

As discussed in the reviews [68,69], other researchers have developed a variety of methods to automatically explain the predictions made by machine learning models. Many of these methods lower the model performance or work only for a specific machine learning algorithm. Most of these methods provide explanations that are not of rule types. More importantly, none of these methods can automatically suggest tailored interventions, which is desired in many clinical applications. In comparison, our automatic explanation method has four properties that make it particularly suitable for providing clinical decision support: (1) it provides rule-type explanations, which are easier to understand than other kinds of explanations; (2) it works for any machine learning model on tabular data; (3) it does not lower model performance; and

(4) it is the only automatic explanation method that can automatically suggest tailored interventions.

Rudin et al [70], Ribeiro et al [71], Rasouli et al [72], Pastor and Baralis [73], Guidotti et al [74], and Panigutti et al [75] used rules to automatically explain machine learning predictions. These rules are not known before the time of prediction, making it impossible to use them to automatically suggest tailored interventions at the time of prediction. Except for the case of Pastor and Baralis [73], these rules are not association rules. In comparison, our automatic explanation method mines association rules before the time of prediction and uses them to automatically suggest tailored interventions at the time of prediction.

### Limitations

This study has 5 limitations that are worth addressing in future work.

First, this study used data from a single health care system. It is worth assessing our automatic explanation method's performance in explaining the predictions of severe COPD exacerbations in other health care systems.

Second, this study focuses on the prediction of one outcome—whether a patient with COPD will have  $\geq 1$  severe COPD exacerbation in the following 12 months. It is worth assessing our automatic explanation method's performance in explaining the predictions of other outcomes.

Third, our automatic explanation method currently works for explaining the predictions that traditional non-deep-learning machine learning algorithms make on tabular data. It is worth investigating the extension of our method to handle the predictions made by deep learning models on longitudinal data [76,77].

Fourth, we currently know no optimal way to present automatic explanations and automatically suggested interventions. It is worth investigating an optimal way to present this information based on a user-centered design.

Finally, researchers have assessed the impact of automatic explanations on decision-making for several other applications [78-82] before but not for care management. For the automatic explanation function for predicting severe COPD exacerbations presented in this paper, it is worth assessing the impact of showing automatic explanations and automatically suggested interventions on care management enrollment and intervention decisions.

### Conclusions

Our automatic explanation method generalizes well in predicting severe COPD exacerbations. After further improving the performance of our model for predicting severe COPD exacerbations and our automatic explanation method, we hope our model can be used in conjunction with our automatic explanation method to provide decision support for allocating COPD care management resources and improve outcomes.

### Acknowledgments

GL and SZ were partially supported by the National Heart, Lung, and Blood Institute of the National Institutes of Health under award number R01HL142503. SZ was also partially supported by the National Library of Medicine Training Grant under award number T15LM007442. MA was partially supported by grants from the Flight Attendant Medical Research Institute (CIA190001) and the California Tobacco-Related Disease Research Program (T29IR0715). The funders had no role in the study design, data collection and analysis, decision to publish, or preparation of the manuscript.

### Authors' Contributions

GL and SZ were mainly responsible for the paper. SZ conducted a literature review, performed most of the data analysis, and wrote the first draft of the paper. GL conceptualized and designed the study, participated in data analysis, and rewrote the entire paper. MA provided clinical expertise, contributed to conceptualizing the presentation, and revised the paper.

### Conflicts of Interest

None declared.

### References

1. Disease or condition of the week - COPD. Centers for Disease Control and Prevention. 2019. URL: <https://www.cdc.gov/dotw/copd/index.html> [accessed 2021-11-14]
2. Ford ES, Murphy LB, Khavjou O, Giles WH, Holt JB, Croft JB. Total and state-specific medical and absenteeism costs of COPD among adults aged  $\geq 18$  years in the United States for 2010 and projections through 2020. *Chest* 2015 Jan;147(1):31-45. [doi: [10.1378/chest.14-0972](https://doi.org/10.1378/chest.14-0972)] [Medline: [25058738](https://pubmed.ncbi.nlm.nih.gov/25058738/)]
3. 2020 Gold Reports. Global Initiative for Chronic Obstructive Lung Disease - GOLD. 2020. URL: <https://goldcopd.org/gold-reports> [accessed 2021-11-14]
4. Anzueto A, Leimer I, Kesten S. Impact of frequency of COPD exacerbations on pulmonary function, health status and clinical outcomes. *Int J Chron Obstruct Pulmon Dis* 2009;4:245-251 [FREE Full text] [doi: [10.2147/copd.s4862](https://doi.org/10.2147/copd.s4862)] [Medline: [19657398](https://pubmed.ncbi.nlm.nih.gov/19657398/)]

5. Connors Jr AF, Dawson NV, Thomas C, Harrell Jr FE, Desbiens N, Fulkerson WJ, et al. Outcomes following acute exacerbation of severe chronic obstructive lung disease. The SUPPORT investigators (Study to Understand Prognoses and Preferences for Outcomes and Risks of Treatments). *Am J Respir Crit Care Med* 1996 Oct;154(4 Pt 1):959-967. [doi: [10.1164/ajrccm.154.4.8887592](https://doi.org/10.1164/ajrccm.154.4.8887592)] [Medline: [8887592](https://pubmed.ncbi.nlm.nih.gov/8887592/)]
6. Viglio S, Iadarola P, Lupi A, Trisolini R, Tinelli C, Balbi B, et al. MEKC of desmosine and isodesmosine in urine of chronic destructive lung disease patients. *Eur Respir J* 2000 Jun;15(6):1039-1045 [FREE Full text] [doi: [10.1034/j.1399-3003.2000.01511.x](https://doi.org/10.1034/j.1399-3003.2000.01511.x)] [Medline: [10885422](https://pubmed.ncbi.nlm.nih.gov/10885422/)]
7. Kanner RE, Anthonisen NR, Connett JE, Lung Health Study Research Group. Lower respiratory illnesses promote FEV(1) decline in current smokers but not ex-smokers with mild chronic obstructive pulmonary disease: results from the lung health study. *Am J Respir Crit Care Med* 2001 Aug 01;164(3):358-364. [doi: [10.1164/ajrccm.164.3.2010017](https://doi.org/10.1164/ajrccm.164.3.2010017)] [Medline: [11500333](https://pubmed.ncbi.nlm.nih.gov/11500333/)]
8. Spencer S, Jones PW, GLOBE Study Group. Time course of recovery of health status following an infective exacerbation of chronic bronchitis. *Thorax* 2003 Jul;58(7):589-593 [FREE Full text] [doi: [10.1136/thorax.58.7.589](https://doi.org/10.1136/thorax.58.7.589)] [Medline: [12832673](https://pubmed.ncbi.nlm.nih.gov/12832673/)]
9. Spencer S, Calverley PM, Burge PS, Jones PW, ISOLDE (Inhaled Steroids in Obstructive Lung Disease) Study Group. Health status deterioration in patients with chronic obstructive pulmonary disease. *Am J Respir Crit Care Med* 2001 Jan;163(1):122-128. [doi: [10.1164/ajrccm.163.1.2005009](https://doi.org/10.1164/ajrccm.163.1.2005009)] [Medline: [11208636](https://pubmed.ncbi.nlm.nih.gov/11208636/)]
10. Blanchette CM, Dalal AA, Mapel D. Changes in COPD demographics and costs over 20 years. *J Med Econ* 2012;15(6):1176-1182. [doi: [10.3111/13696998.2012.713880](https://doi.org/10.3111/13696998.2012.713880)] [Medline: [22812689](https://pubmed.ncbi.nlm.nih.gov/22812689/)]
11. Johnston J, Longman J, Ewald D, King J, Das S, Passey M. Study of potentially preventable hospitalisations (PPH) for chronic conditions: what proportion are preventable and what factors are associated with preventable PPH? *BMJ Open* 2020 Nov 09;10(11):e038415 [FREE Full text] [doi: [10.1136/bmjopen-2020-038415](https://doi.org/10.1136/bmjopen-2020-038415)] [Medline: [33168551](https://pubmed.ncbi.nlm.nih.gov/33168551/)]
12. Curry N, Billings J, Darin B, Dixon J, Williams M, Wennberg D. Predictive risk project literature review. King's Fund, London. 2005. URL: [http://www.kingsfund.org.uk/sites/files/kf/field/field\\_document/predictive-risk-literature-review-june2005.pdf](http://www.kingsfund.org.uk/sites/files/kf/field/field_document/predictive-risk-literature-review-june2005.pdf). [accessed 2021-11-14]
13. Mays GP, Claxton G, White J. Managed care rebound? Recent changes in health plans' cost containment strategies. *Health Aff (Millwood)* 2004;Suppl Web Exclusives:427-436 [FREE Full text] [doi: [10.1377/hlthaff.w4.427](https://doi.org/10.1377/hlthaff.w4.427)] [Medline: [15451964](https://pubmed.ncbi.nlm.nih.gov/15451964/)]
14. Bandurska E, Damps-Konstańska I, Popowski P, Jędrzejczyk T, Janowiak P, Świętnicka K, et al. Impact of integrated care model (ICM) on direct medical costs in management of advanced chronic obstructive pulmonary disease (COPD). *Med Sci Monit* 2017 Jun 12;23:2850-2862 [FREE Full text] [doi: [10.12659/msm.901982](https://doi.org/10.12659/msm.901982)] [Medline: [28603270](https://pubmed.ncbi.nlm.nih.gov/28603270/)]
15. Rice KL, Dewan N, Bloomfield HE, Grill J, Schult TM, Nelson DB, et al. Disease management program for chronic obstructive pulmonary disease: a randomized controlled trial. *Am J Respir Crit Care Med* 2010 Oct 1;182(7):890-896. [doi: [10.1164/rccm.200910-1579OC](https://doi.org/10.1164/rccm.200910-1579OC)] [Medline: [20075385](https://pubmed.ncbi.nlm.nih.gov/20075385/)]
16. Axelrod RC, Vogel D. Predictive modeling in health plans. *Dis Manag Health Outcomes* 2003;11(12):779-787. [doi: [10.2165/00115677-200311120-00003](https://doi.org/10.2165/00115677-200311120-00003)]
17. Zeng S, Arjomandi M, Tong Y, Liao ZC, Luo G. Developing a machine learning model to predict severe chronic obstructive pulmonary disease exacerbations: retrospective cohort study. *J Med Internet Res* 2022 Jan 06;24(1):e28953 [FREE Full text] [doi: [10.2196/28953](https://doi.org/10.2196/28953)] [Medline: [34989686](https://pubmed.ncbi.nlm.nih.gov/34989686/)]
18. Annavarapu S, Goldfarb S, Gelb M, Moretz C, Renda A, Kaila S. Development and validation of a predictive model to identify patients at risk of severe COPD exacerbations using administrative claims data. *Int J Chron Obstruct Pulmon Dis* 2018;13:2121-2130 [FREE Full text] [doi: [10.2147/COPD.S155773](https://doi.org/10.2147/COPD.S155773)] [Medline: [30022818](https://pubmed.ncbi.nlm.nih.gov/30022818/)]
19. Tavakoli H, Chen W, Sin DD, FitzGerald JM, Sadatsafavi M. Predicting severe chronic obstructive pulmonary disease exacerbations. Developing a population surveillance approach with administrative data. *Ann Am Thorac Soc* 2020 Sep;17(9):1069-1076. [doi: [10.1513/AnnalsATS.202001-070OC](https://doi.org/10.1513/AnnalsATS.202001-070OC)] [Medline: [32383971](https://pubmed.ncbi.nlm.nih.gov/32383971/)]
20. Samp JC, Joo MJ, Schumock GT, Calip GS, Pickard AS, Lee TA. Predicting acute exacerbations in chronic obstructive pulmonary disease. *J Manag Care Spec Pharm* 2018 Mar;24(3):265-279. [doi: [10.18553/jmcp.2018.24.3.265](https://doi.org/10.18553/jmcp.2018.24.3.265)] [Medline: [29485951](https://pubmed.ncbi.nlm.nih.gov/29485951/)]
21. Thomsen M, Ingebrigtsen TS, Marott JL, Dahl M, Lange P, Vestbo J, et al. Inflammatory biomarkers and exacerbations in chronic obstructive pulmonary disease. *J Am Med Assoc* 2013 Jun 12;309(22):2353-2361. [doi: [10.1001/jama.2013.5732](https://doi.org/10.1001/jama.2013.5732)] [Medline: [23757083](https://pubmed.ncbi.nlm.nih.gov/23757083/)]
22. Orchard P, Agakova A, Pinnock H, Burton CD, Sarran C, Agakov F, et al. Improving prediction of risk of hospital admission in chronic obstructive pulmonary disease: application of machine learning to telemonitoring data. *J Med Internet Res* 2018 Sep 21;20(9):e263 [FREE Full text] [doi: [10.2196/jmir.9227](https://doi.org/10.2196/jmir.9227)] [Medline: [30249589](https://pubmed.ncbi.nlm.nih.gov/30249589/)]
23. Suetomo M, Kawayama T, Kinoshita T, Takenaka S, Matsuoka M, Matsunaga K, et al. COPD assessment tests scores are associated with exacerbated chronic obstructive pulmonary disease in Japanese patients. *Respir Investig* 2014 Sep;52(5):288-295. [doi: [10.1016/j.resinv.2014.04.004](https://doi.org/10.1016/j.resinv.2014.04.004)] [Medline: [25169844](https://pubmed.ncbi.nlm.nih.gov/25169844/)]
24. Lee SD, Huang MS, Kang J, Lin CH, Park MJ, Oh YM, Investigators of the Predictive Ability of CAT in Acute Exacerbations of COPD (PACE) Study. The COPD assessment test (CAT) assists prediction of COPD exacerbations in high-risk patients. *Respir Med* 2014 Apr;108(4):600-608 [FREE Full text] [doi: [10.1016/j.rmed.2013.12.014](https://doi.org/10.1016/j.rmed.2013.12.014)] [Medline: [24456695](https://pubmed.ncbi.nlm.nih.gov/24456695/)]



25. Faganello MM, Tanni SE, Sanchez FF, Pelegrino NR, Lucheta PA, Godoy I. BODE index and GOLD staging as predictors of 1-year exacerbation risk in chronic obstructive pulmonary disease. *Am J Med Sci* 2010 Jan;339(1):10-14. [doi: [10.1097/MAJ.0b013e3181bb8111](https://doi.org/10.1097/MAJ.0b013e3181bb8111)] [Medline: [19926966](https://pubmed.ncbi.nlm.nih.gov/19926966/)]
26. Alcázar B, García-Polo C, Herrejón A, Ruiz LA, de Miguel J, Ros JA, et al. Factors associated with hospital admission for exacerbation of chronic obstructive pulmonary disease. *Arch Bronconeumol* 2012 Mar;48(3):70-76. [doi: [10.1016/j.arbres.2011.10.009](https://doi.org/10.1016/j.arbres.2011.10.009)] [Medline: [22196478](https://pubmed.ncbi.nlm.nih.gov/22196478/)]
27. Bertens LC, Reitsma JB, Moons KG, van Mourik Y, Lammers JW, Broekhuizen BD, et al. Development and validation of a model to predict the risk of exacerbations in chronic obstructive pulmonary disease. *Int J Chron Obstruct Pulmon Dis* 2013;8:493-499 [FREE Full text] [doi: [10.2147/COPD.S49609](https://doi.org/10.2147/COPD.S49609)] [Medline: [24143086](https://pubmed.ncbi.nlm.nih.gov/24143086/)]
28. Miravittles M, Guerrero T, Mayordomo C, Sánchez-Agudo L, Nicolau F, Segú JL. Factors associated with increased risk of exacerbation and hospital admission in a cohort of ambulatory COPD patients: a multiple logistic regression analysis. The EOLO Study Group. *Respiration* 2000;67(5):495-501. [doi: [10.1159/000067462](https://doi.org/10.1159/000067462)] [Medline: [11070451](https://pubmed.ncbi.nlm.nih.gov/11070451/)]
29. Make BJ, Eriksson G, Calverley PM, Jenkins CR, Postma DS, Peterson S, et al. A score to predict short-term risk of COPD exacerbations (SCOPEX). *Int J Chron Obstruct Pulmon Dis* 2015;10:201-209 [FREE Full text] [doi: [10.2147/COPD.S69589](https://doi.org/10.2147/COPD.S69589)] [Medline: [25670896](https://pubmed.ncbi.nlm.nih.gov/25670896/)]
30. Montserrat-Capdevila J, Godoy P, Marsal JR, Barbé F. Predictive model of hospital admission for COPD exacerbation. *Respir Care* 2015 Sep;60(9):1288-1294 [FREE Full text] [doi: [10.4187/respcare.04005](https://doi.org/10.4187/respcare.04005)] [Medline: [26286737](https://pubmed.ncbi.nlm.nih.gov/26286737/)]
31. Kerkhof M, Freeman D, Jones R, Chisholm A, Price DB, Respiratory Effectiveness Group. Predicting frequent COPD exacerbations using primary care data. *Int J Chron Obstruct Pulmon Dis* 2015;10:2439-2450 [FREE Full text] [doi: [10.2147/COPD.S94259](https://doi.org/10.2147/COPD.S94259)] [Medline: [26609229](https://pubmed.ncbi.nlm.nih.gov/26609229/)]
32. Chen X, Wang Q, Hu Y, Zhang L, Xiong W, Xu Y, et al. A nomogram for predicting severe exacerbations in stable COPD patients. *Int J Chron Obstruct Pulmon Dis* 2020;15:379-388 [FREE Full text] [doi: [10.2147/COPD.S234241](https://doi.org/10.2147/COPD.S234241)] [Medline: [32110006](https://pubmed.ncbi.nlm.nih.gov/32110006/)]
33. Yii AC, Loh CH, Tiew PY, Xu H, Taha AA, Koh J, et al. A clinical prediction model for hospitalized COPD exacerbations based on "treatable traits". *Int J Chron Obstruct Pulmon Dis* 2019;14:719-728 [FREE Full text] [doi: [10.2147/COPD.S194922](https://doi.org/10.2147/COPD.S194922)] [Medline: [30988606](https://pubmed.ncbi.nlm.nih.gov/30988606/)]
34. 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](https://doi.org/10.1016/S2213-2600(19)30397-2)] [Medline: [32178776](https://pubmed.ncbi.nlm.nih.gov/32178776/)]
35. Stanford RH, Nag A, Mapel DW, Lee TA, Rosiello R, Vekeman F, et al. Validation of a new risk measure for chronic obstructive pulmonary disease exacerbation using health insurance claims data. *Ann Am Thorac Soc* 2016 Jul;13(7):1067-1075. [doi: [10.1513/AnnalsATS.201508-493OC](https://doi.org/10.1513/AnnalsATS.201508-493OC)] [Medline: [27070274](https://pubmed.ncbi.nlm.nih.gov/27070274/)]
36. Stanford RH, Nag A, Mapel DW, Lee TA, Rosiello R, Schatz M, et al. Claims-based risk model for first severe COPD exacerbation. *Am J Manag Care* 2018 Feb 01;24(2):45-53 [FREE Full text] [Medline: [29461849](https://pubmed.ncbi.nlm.nih.gov/29461849/)]
37. Stanford RH, Lau MS, Li Y, Stemkowski S. External validation of a COPD risk measure in a commercial and Medicare population: the COPD treatment ratio. *J Manag Care Spec Pharm* 2019 Jan;25(1):58-69. [doi: [10.18553/jmcp.2019.25.1.058](https://doi.org/10.18553/jmcp.2019.25.1.058)] [Medline: [30589629](https://pubmed.ncbi.nlm.nih.gov/30589629/)]
38. Stanford RH, Korrer S, Brekke L, Reinsch T, Bengtson LG. Validation and assessment of the COPD treatment ratio as a predictor of severe exacerbations. *Chronic Obstr Pulm Dis* 2020 Jan;7(1):38-48 [FREE Full text] [doi: [10.15326/jcopdf.7.1.2019.0132](https://doi.org/10.15326/jcopdf.7.1.2019.0132)] [Medline: [31999901](https://pubmed.ncbi.nlm.nih.gov/31999901/)]
39. Jones RC, Donaldson GC, Chavannes NH, Kida K, Dickson-Spillmann M, Harding S, et al. Derivation and validation of a composite index of severity in chronic obstructive pulmonary disease: the DOSE Index. *Am J Respir Crit Care Med* 2009 Dec 15;180(12):1189-1195. [doi: [10.1164/rccm.200902-0271OC](https://doi.org/10.1164/rccm.200902-0271OC)] [Medline: [19797160](https://pubmed.ncbi.nlm.nih.gov/19797160/)]
40. Jones RC, Price D, Chavannes NH, Lee AJ, Hyland ME, Ställberg B, UNLOCK Group of the IPCRG. Multi-component assessment of chronic obstructive pulmonary disease: an evaluation of the ADO and DOSE indices and the global obstructive lung disease categories in international primary care data sets. *NPJ Prim Care Respir Med* 2016 Apr 07;26:16010 [FREE Full text] [doi: [10.1038/npjpcrm.2016.10](https://doi.org/10.1038/npjpcrm.2016.10)] [Medline: [27053297](https://pubmed.ncbi.nlm.nih.gov/27053297/)]
41. Fan VS, Curtis JR, Tu SP, McDonnell MB, Fihn SD, Ambulatory Care Quality Improvement Project Investigators. Using quality of life to predict hospitalization and mortality in patients with obstructive lung diseases. *Chest* 2002 Aug;122(2):429-436. [doi: [10.1378/chest.122.2.429](https://doi.org/10.1378/chest.122.2.429)] [Medline: [12171813](https://pubmed.ncbi.nlm.nih.gov/12171813/)]
42. Moy ML, Teylan M, Danilack VA, Gagnon DR, Garshick E. An index of daily step count and systemic inflammation predicts clinical outcomes in chronic obstructive pulmonary disease. *Ann Am Thorac Soc* 2014 Feb;11(2):149-157. [doi: [10.1513/AnnalsATS.201307-243OC](https://doi.org/10.1513/AnnalsATS.201307-243OC)] [Medline: [24308588](https://pubmed.ncbi.nlm.nih.gov/24308588/)]
43. Briggs A, Spencer M, Wang H, Mannino D, Sin DD. Development and validation of a prognostic index for health outcomes in chronic obstructive pulmonary disease. *Arch Intern Med* 2008 Jan 14;168(1):71-79. [doi: [10.1001/archinternmed.2007.37](https://doi.org/10.1001/archinternmed.2007.37)] [Medline: [18195198](https://pubmed.ncbi.nlm.nih.gov/18195198/)]
44. Lange P, Marott JL, Vestbo J, Olsen KR, Ingebrigtsen TS, Dahl M, et al. Prediction of the clinical course of chronic obstructive pulmonary disease, using the new GOLD classification: a study of the general population. *Am J Respir Crit Care Med* 2012 Nov 15;186(10):975-981. [doi: [10.1164/rccm.201207-1299OC](https://doi.org/10.1164/rccm.201207-1299OC)] [Medline: [22997207](https://pubmed.ncbi.nlm.nih.gov/22997207/)]

45. Abascal-Bolado B, Novotny PJ, Sloan JA, Karpman C, Dulohery MM, Benzo RP. Forecasting COPD hospitalization in the clinic: optimizing the chronic respiratory questionnaire. *Int J Chron Obstruct Pulmon Dis* 2015;10:2295-2301 [FREE Full text] [doi: [10.2147/COPD.S87469](https://doi.org/10.2147/COPD.S87469)] [Medline: [26543362](https://pubmed.ncbi.nlm.nih.gov/26543362/)]
46. Blanco-Aparicio M, Vázquez I, Pita-Fernández S, Pértega-Díaz S, Vereá-Hernando H. Utility of brief questionnaires of health-related quality of life (Airways Questionnaire 20 and Clinical COPD Questionnaire) to predict exacerbations in patients with asthma and COPD. *Health Qual Life Outcomes* 2013 May 27;11:85 [FREE Full text] [doi: [10.1186/1477-7525-11-85](https://doi.org/10.1186/1477-7525-11-85)] [Medline: [23706146](https://pubmed.ncbi.nlm.nih.gov/23706146/)]
47. Yoo JW, Hong Y, Seo JB, Chae EJ, Ra SW, Lee JH, et al. Comparison of clinico-physiologic and CT imaging risk factors for COPD exacerbation. *J Korean Med Sci* 2011 Dec;26(12):1606-1612 [FREE Full text] [doi: [10.3346/jkms.2011.26.12.1606](https://doi.org/10.3346/jkms.2011.26.12.1606)] [Medline: [22147998](https://pubmed.ncbi.nlm.nih.gov/22147998/)]
48. Niewoehner DE, Likhnygina Y, Rice K, Kuschner WG, Sharafkhaneh A, Sarosi GA, et al. Risk indexes for exacerbations and hospitalizations due to COPD. *Chest* 2007 Jan;131(1):20-28. [doi: [10.1378/chest.06-1316](https://doi.org/10.1378/chest.06-1316)] [Medline: [17218552](https://pubmed.ncbi.nlm.nih.gov/17218552/)]
49. Austin PC, Stanbrook MB, Anderson GM, Newman A, Gershon AS. Comparative ability of comorbidity classification methods for administrative data to predict outcomes in patients with chronic obstructive pulmonary disease. *Ann Epidemiol* 2012 Dec;22(12):881-887 [FREE Full text] [doi: [10.1016/j.annepidem.2012.09.011](https://doi.org/10.1016/j.annepidem.2012.09.011)] [Medline: [23121992](https://pubmed.ncbi.nlm.nih.gov/23121992/)]
50. Marin JM, Carrizo SJ, Casanova C, Martínez-Cambor P, Soriano JB, Agustí AG, et al. Prediction of risk of COPD exacerbations by the BODE index. *Respir Med* 2009 Mar;103(3):373-378 [FREE Full text] [doi: [10.1016/j.rmed.2008.10.004](https://doi.org/10.1016/j.rmed.2008.10.004)] [Medline: [19013781](https://pubmed.ncbi.nlm.nih.gov/19013781/)]
51. Ställberg B, Lisspers K, Larsson K, Janson C, Müller M, Łuczko M, et al. Predicting hospitalization due to COPD exacerbations in Swedish primary care patients using machine learning - based on the ARCTIC study. *Int J Chron Obstruct Pulmon Dis* 2021;16:677-688 [FREE Full text] [doi: [10.2147/COPD.S293099](https://doi.org/10.2147/COPD.S293099)] [Medline: [33758504](https://pubmed.ncbi.nlm.nih.gov/33758504/)]
52. Chen T, Guestrin C. XGBoost: a scalable tree boosting system. In: *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. 2016 Presented at: KDD'16: The 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining; August 13-17, 2016; San Francisco, CA p. 785-794. [doi: [10.1145/2939672.2939785](https://doi.org/10.1145/2939672.2939785)]
53. U.S. healthcare leaders expect widespread adoption of artificial intelligence by 2023. Intel. 2018. URL: <https://newsroom.intel.com/news-releases/u-s-healthcare-leaders-expect-widespread-adoption-artificial-intelligence-2023> [accessed 2021-11-14]
54. Luo G, Johnson MD, Nkoy FL, He S, Stone BL. Automatically explaining machine learning prediction results on asthma hospital visits in patients with asthma: secondary analysis. *JMIR Med Inform* 2020 Dec 31;8(12):e21965 [FREE Full text] [doi: [10.2196/21965](https://doi.org/10.2196/21965)] [Medline: [33382379](https://pubmed.ncbi.nlm.nih.gov/33382379/)]
55. Tong Y, Messinger AI, Luo G. Testing the generalizability of an automated method for explaining machine learning predictions on asthma patients' asthma hospital visits to an academic healthcare system. *IEEE Access* 2020;8:195971-195979 [FREE Full text] [doi: [10.1109/access.2020.3032683](https://doi.org/10.1109/access.2020.3032683)] [Medline: [33240737](https://pubmed.ncbi.nlm.nih.gov/33240737/)]
56. Luo G, Nau CL, Crawford WW, Schatz M, Zeiger RS, Koebnick C. Generalizability of an automatic explanation method for machine learning prediction results on asthma-related hospital visits in patients with asthma: quantitative analysis. *J Med Internet Res* 2021 Apr 15;23(4):e24153 [FREE Full text] [doi: [10.2196/24153](https://doi.org/10.2196/24153)] [Medline: [33856359](https://pubmed.ncbi.nlm.nih.gov/33856359/)]
57. Zhang X, Luo G. Ranking rule-based automatic explanations for machine learning predictions on asthma hospital encounters in patients with asthma: retrospective cohort study. *JMIR Med Inform* 2021 Aug 11;9(8):e28287 [FREE Full text] [doi: [10.2196/28287](https://doi.org/10.2196/28287)] [Medline: [34383673](https://pubmed.ncbi.nlm.nih.gov/34383673/)]
58. Luo G. Automatically explaining machine learning prediction results: a demonstration on type 2 diabetes risk prediction. *Health Inf Sci Syst* 2016;4:2 [FREE Full text] [doi: [10.1186/s13755-016-0015-4](https://doi.org/10.1186/s13755-016-0015-4)] [Medline: [26958341](https://pubmed.ncbi.nlm.nih.gov/26958341/)]
59. NQF #1891 Hospital 30-day, all-cause, risk-standardized readmission rate (RSRR) following chronic obstructive pulmonary disease (COPD) hospitalization. National Quality Forum. 2012. URL: [http://www.qualityforum.org/Projects/n-r/Pulmonary\\_Endorsement\\_Maintenance/1891\\_30\\_Day\\_RSRR\\_COPD.aspx](http://www.qualityforum.org/Projects/n-r/Pulmonary_Endorsement_Maintenance/1891_30_Day_RSRR_COPD.aspx) [accessed 2021-11-14]
60. Cooke CR, Joo MJ, Anderson SM, Lee TA, Udriș EM, Johnson E, et al. The validity of using ICD-9 codes and pharmacy records to identify patients with chronic obstructive pulmonary disease. *BMC Health Serv Res* 2011 Feb 16;11:37 [FREE Full text] [doi: [10.1186/1472-6963-11-37](https://doi.org/10.1186/1472-6963-11-37)] [Medline: [21324188](https://pubmed.ncbi.nlm.nih.gov/21324188/)]
61. Nguyen HQ, Chu L, Amy Liu LI, Lee JS, Suh D, Korotzer B, et al. Associations between physical activity and 30-day readmission risk in chronic obstructive pulmonary disease. *Ann Am Thorac Soc* 2014 Jun;11(5):695-705. [doi: [10.1513/AnnalsATS.201401-017OC](https://doi.org/10.1513/AnnalsATS.201401-017OC)] [Medline: [24713094](https://pubmed.ncbi.nlm.nih.gov/24713094/)]
62. Lindenauer PK, Grosso LM, Wang C, Wang Y, Krishnan JA, Lee TA, et al. Development, validation, and results of a risk-standardized measure of hospital 30-day mortality for patients with exacerbation of chronic obstructive pulmonary disease. *J Hosp Med* 2013 Aug;8(8):428-435. [doi: [10.1002/jhm.2066](https://doi.org/10.1002/jhm.2066)] [Medline: [23913593](https://pubmed.ncbi.nlm.nih.gov/23913593/)]
63. Liu B, Hsu W, Ma Y. Integrating classification and association rule mining. In: *Proceedings of the 4th International Conference on Knowledge Discovery and Data Mining*. 1998 Presented at: KDD'98: Fourth International Conference on Knowledge Discovery and Data Mining; August 27-31, 1998; New York City, NY, USA p. 80-86 URL: <https://dl.acm.org/doi/10.5555/3000292.3000305>
64. Thabtah FA. A review of associative classification mining. *Knowledge Eng Review* 2007 Mar 01;22(1):37-65. [doi: [10.1017/s026988907001026](https://doi.org/10.1017/s026988907001026)]

65. Fayyad UM, Irani KB. Multi-interval discretization of continuous-valued attributes for classification learning. In: Proceedings of the 13th International Joint Conference on Artificial Intelligence. 1993 Presented at: IJCAI'93: 13th International Joint Conference on Artificial Intelligence; August 28-September 3, 1993; Chambéry, France p. 1022-1029 URL: <https://trs.jpl.nasa.gov/handle/2014/35171>
66. Alaa AM, van der Schaar M. Prognostication and risk factors for cystic fibrosis via automated machine learning. *Sci Rep* 2018 Jul 26;8(1):11242 [FREE Full text] [doi: [10.1038/s41598-018-29523-2](https://doi.org/10.1038/s41598-018-29523-2)] [Medline: [30050169](https://pubmed.ncbi.nlm.nih.gov/30050169/)]
67. Alaa AM, van der Schaar M. AutoPrognosis: automated clinical prognostic modeling via Bayesian optimization with structured kernel learning. In: Proceedings of the 35th International Conference on Machine Learning. 2018 Presented at: ICML'18: 35th International Conference on Machine Learning; July 10-15, 2018; Stockholm, Sweden p. 139-148 URL: <https://arxiv.org/abs/1802.07207>
68. Guidotti R, Monreale A, Ruggieri S, Turini F, Giannotti F, Pedreschi D. A survey of methods for explaining black box models. *ACM Comput Surv* 2019 Sep 30;51(5):93. [doi: [10.1145/3236009](https://doi.org/10.1145/3236009)]
69. Payrovnaziri SN, Chen Z, Rengifo-Moreno P, Miller T, Bian J, Chen JH, et al. Explainable artificial intelligence models using real-world electronic health record data: a systematic scoping review. *J Am Med Inform Assoc* 2020 Jul 01;27(7):1173-1185 [FREE Full text] [doi: [10.1093/jamia/ocaa053](https://doi.org/10.1093/jamia/ocaa053)] [Medline: [32417928](https://pubmed.ncbi.nlm.nih.gov/32417928/)]
70. Rudin C, Shaposhnik Y. Globally-consistent rule-based summary-explanations for machine learning models: application to credit-risk evaluation. In: Proceedings of INFORMS 11th Conference on Information Systems and Technology. 2019 Presented at: CIST'19: INFORMS 11th Conference on Information Systems and Technology; October 19-20, 2019; Seattle, WA p. 1-19. [doi: [10.2139/ssrn.3395422](https://doi.org/10.2139/ssrn.3395422)]
71. Ribeiro MT, Singh S, Guestrin C. Anchors: high-precision model-agnostic explanations. In: Proceedings of the 32nd AAAI Conference on Artificial Intelligence. 2018 Presented at: AAAI'18: 32nd AAAI Conference on Artificial Intelligence; February 2-7, 2018; New Orleans, LA p. 1527-1535 URL: <https://ojs.aaai.org/index.php/AAAI/article/view/11491>
72. Rasouli P, Yu IC. EXPLAN: explaining black-box classifiers using adaptive neighborhood generation. In: Proceedings of the 2020 International Joint Conference on Neural Networks. 2020 Presented at: IJCNN'20: International Joint Conference on Neural Networks; July 19-24, 2020; Glasgow, UK p. 1-9. [doi: [10.1109/ijcnn48605.2020.9206710](https://doi.org/10.1109/ijcnn48605.2020.9206710)]
73. Pastor E, Baralis E. Explaining black box models by means of local rules. In: Proceedings of the 34th ACM/SIGAPP Symposium on Applied Computing. 2019 Presented at: SAC'19: The 34th ACM/SIGAPP Symposium on Applied Computing; April 8-12, 2019; Limassol, Cyprus p. 510-517. [doi: [10.1145/3297280.3297328](https://doi.org/10.1145/3297280.3297328)]
74. Guidotti R, Monreale A, Ruggieri S, Pedreschi D, Turini F, Giannotti F. Local rule-based explanations of black box decision systems. *arXiv*. 2018. URL: <https://arxiv.org/abs/1805.10820> [accessed 2021-11-14]
75. Panigutti C, Perotti A, Pedreschi D. Doctor XAI: an ontology-based approach to black-box sequential data classification explanations. In: Proceedings of the Conference on Fairness, Accountability, and Transparency. 2020 Presented at: FAT'20: Conference on Fairness, Accountability, and Transparency; January 27-30, 2020; Barcelona, Spain p. 629-639. [doi: [10.1145/3351095.3372855](https://doi.org/10.1145/3351095.3372855)]
76. Luo G, Stone BL, Koebnick C, He S, Au DH, Sheng X, et al. Using temporal features to provide data-driven clinical early warnings for chronic obstructive pulmonary disease and asthma care management: protocol for a secondary analysis. *JMIR Res Protoc* 2019 Jun 06;8(6):e13783 [FREE Full text] [doi: [10.2196/13783](https://doi.org/10.2196/13783)] [Medline: [31199308](https://pubmed.ncbi.nlm.nih.gov/31199308/)]
77. Luo G. A roadmap for semi-automatically extracting predictive and clinically meaningful temporal features from medical data for predictive modeling. *Glob Transit* 2019;1:61-82 [FREE Full text] [doi: [10.1016/j.glt.2018.11.001](https://doi.org/10.1016/j.glt.2018.11.001)] [Medline: [31032483](https://pubmed.ncbi.nlm.nih.gov/31032483/)]
78. Weerts HJ, van Ipenburg W, Pechenizkiy M. A human-grounded evaluation of SHAP for alert processing. *arXiv*. 2019. URL: <https://arxiv.org/abs/1907.03324> [accessed 2021-11-14]
79. Stites MC, Nyre-Yu M, Moss B, Smutz C, Smith MR. Sage advice? The impacts of explanations for machine learning models on human decision-making in spam detection. In: Proceedings of the Second International Conference on Artificial Intelligence in HCI. 2021 Presented at: AI-HCI'21: International Conference on Artificial Intelligence in HCI; July 24-29, 2021; Washington DC p. 269-284. [doi: [10.1007/978-3-030-77772-2\\_18](https://doi.org/10.1007/978-3-030-77772-2_18)]
80. Lai V, Tan C. On human predictions with explanations and predictions of machine learning models: a case study on deception detection. In: Proceedings of the Conference on Fairness, Accountability, and Transparency. 2019 Presented at: FAT'19: Conference on Fairness, Accountability, and Transparency; January 29-31, 2019; Atlanta, GA p. 29-38. [doi: [10.1145/3287560.3287590](https://doi.org/10.1145/3287560.3287590)]
81. 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](https://doi.org/10.1038/s41551-018-0304-0)] [Medline: [31001455](https://pubmed.ncbi.nlm.nih.gov/31001455/)]
82. Jesus SM, Belém C, Balayan V, Bento J, Saleiro P, Bizarro P, et al. How can I choose an explainer? An application-grounded evaluation of post-hoc explanations. In: Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency. 2021 Presented at: FAccT'21: 2021 ACM Conference on Fairness, Accountability, and Transparency; March 3-10, 2021; Toronto, Canada p. 805-815. [doi: [10.1145/3442188.3445941](https://doi.org/10.1145/3442188.3445941)]

## Abbreviations

**COPD:** chronic obstructive pulmonary disease

**ED:** emergency department

**UWM:** University of Washington Medicine

**XGBoost:** extreme gradient boosting

*Edited by C Lovis; submitted 26.08.21; peer-reviewed by P Orchard, A Rovetta; comments to author 13.11.21; revised version received 15.11.21; accepted 02.01.22; published 25.02.22.*

*Please cite as:*

Zeng S, Arjomandi M, Luo G

*Automatically Explaining Machine Learning Predictions on Severe Chronic Obstructive Pulmonary Disease Exacerbations: Retrospective Cohort Study*

*JMIR Med Inform 2022;10(2):e33043*

URL: <https://medinform.jmir.org/2022/2/e33043>

doi: [10.2196/33043](https://doi.org/10.2196/33043)

PMID: [35212634](https://pubmed.ncbi.nlm.nih.gov/35212634/)

©Siyang Zeng, Mehrdad Arjomandi, Gang Luo. Originally published in JMIR Medical Informatics (<https://medinform.jmir.org>), 25.02.2022. This is an open-access article distributed under the terms of the Creative Commons Attribution License (<https://creativecommons.org/licenses/by/4.0/>), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Medical Informatics, is properly cited. The complete bibliographic information, a link to the original publication on <https://medinform.jmir.org/>, as well as this copyright and license information must be included.

Original Paper

# Exploring Quality Differences in Telemedicine Between Hospital Outpatient Departments and Community Clinics: Cross-sectional Study

Noora Alhajri<sup>1</sup>, MD, MPH; Mecit Can Emre Simsekler<sup>2</sup>, PhD; Buthaina Alfalasi<sup>3</sup>, MD; Mohamed Alhashmi<sup>1</sup>, BEng; Hamda Memon<sup>1</sup>, BSc; Emma Housser<sup>1</sup>, MSc; Abdulhamid Mustafa Abdi<sup>1</sup>, BSc; Nahed Balalaa<sup>4</sup>, MD; Maryam Al Ali<sup>5</sup>, MD; Raghdha Almaashari<sup>6</sup>, MD; Shammah Al Memari<sup>7</sup>, MBBS; Farida Al Hosani<sup>7</sup>, MD; Yousif Al Zaabi<sup>8</sup>, BDS, MS; Shereena Almazrouei<sup>7</sup>, MD; Hamed Alhashemi<sup>8</sup>, PhD

<sup>1</sup>College of Medicine and Health Science, Khalifa University, Abu Dhabi, United Arab Emirates

<sup>2</sup>College of Engineering, Khalifa University, Department of Industrial and Systems Engineering, Abu Dhabi, United Arab Emirates

<sup>3</sup>Department of Family Medicine, Zayed Military Hospital, Abu Dhabi, United Arab Emirates

<sup>4</sup>Department of General Surgery, Sheikh Shakhbout Medical City, Abu Dhabi, United Arab Emirates

<sup>5</sup>Ambulatory Healthcare Services, Abu Dhabi Health Services Company, Abu Dhabi, United Arab Emirates

<sup>6</sup>Department of Dermatology, Sheikh Khalifa Medical City, Abu Dhabi, United Arab Emirates

<sup>7</sup>Abu Dhabi Public Health Center, Abu Dhabi, United Arab Emirates

<sup>8</sup>Department of Health, Abu Dhabi, United Arab Emirates

**Corresponding Author:**

Noora Alhajri, MD, MPH

College of Medicine and Health Science

Khalifa University

Al-Saada Road, Zone 1

P O Box 11001

Abu Dhabi

United Arab Emirates

Phone: 971 552277490

Email: [nalhajri007@gmail.com](mailto:nalhajri007@gmail.com)

## Abstract

**Background:** Telemedicine is a care delivery modality that has the potential to broaden the reach and flexibility of health care services. In the United Arab Emirates, telemedicine services are mainly delivered through either integrated hospital outpatient department (OPDs) or community clinics. However, it is unknown if patients' perceptions of, and satisfaction with, telemedicine services differ between these two types of health care systems during the COVID-19 pandemic.

**Objective:** We aimed to explore the differences in patients' perceptions of, and satisfaction with, telemedicine between hospital OPDs and community clinics during the COVID-19 pandemic. We also aimed to identify patient- or visit-related characteristics contributing to patient satisfaction with telemedicine.

**Methods:** In this cross-sectional study that was conducted at Abu Dhabi health care centers, we invited outpatients aged 18 years or over, who completed a telemedicine visit during the COVID-19 pandemic, to participate in our study. Patients' perceptions of, and satisfaction with, telemedicine regarding the two system types (ie, hospital OPDs and community clinics) were assessed using an online survey that was sent as a link through the SMS system. Regression models were used to describe the association between patient- and visit-related characteristics, as well as the perception of, and satisfaction with, telemedicine services.

**Results:** A total of 515 patients participated in this survey. Patients' satisfaction with telemedicine services was equally high among the settings, with no statistically significant difference between the two setting types (hospital OPDs: 253/343, 73.8%; community clinics: 114/172, 66.3%;  $P=.19$ ). Video consultation was significantly associated with increased patient satisfaction (odds ratio [OR] 2.57, 95% CI 1.04-6.33;  $P=.04$ ) and patients' support of the transition to telemedicine use during and after the pandemic (OR 2.88, 95% CI 1.18-7.07;  $P=.02$ ). Patients who used video consultations were more likely to report that telemedicine improved access to health care services (OR 3.06, 95% CI 1.71-8.03;  $P=.02$ ), reduced waiting times and travel costs (OR 4.94, 95% CI 1.15-21.19;  $P=.03$ ), addressed patients' needs (OR 2.63, 95% CI 1.13-6.11;  $P=.03$ ), and eased expression of patients'

medical concerns during the COVID-19 pandemic (OR 2.19, 95% CI 0.89-5.38;  $P=.09$ ). Surprisingly, middle-aged patients were two times more likely to be satisfied with telemedicine services (OR 2.12, 95% CI 1.09-4.14;  $P=.03$ ), as compared to any other age group in this study.

**Conclusions:** These findings suggest that patient satisfaction was unaffected by the health system setting in which patients received the teleconsultations, whether they were at hospitals or community clinics. Video consultation was associated with increased patient satisfaction with telemedicine services. Efforts should be focused on strategic planning for enhanced telemedicine services, video consultation in particular, for both emergent circumstances, such as the COVID-19 pandemic, and day-to-day health care delivery.

(*JMIR Med Inform* 2022;10(2):e32373) doi:[10.2196/32373](https://doi.org/10.2196/32373)

## KEYWORDS

COVID-19; patient satisfaction; technology acceptance; hospital; community clinic; video consultation; audio consultation; outpatient department; OPD; policy making; UAE

## Introduction

In the face of the COVID-19 pandemic, the world woke up to the limitations of the current health care system [1]. As an analog system, health care was scantily equipped to cope with the rapidly emerging pandemic [2]. The United Arab Emirates (UAE) health care system, like many international health care systems, had been largely based on the “in-person visit” model of care [3]. This care delivery model was challenging during the COVID-19 pandemic, given the fast spread of the virus and risk of transmission to uninfected patients who were seeking medical assessment [2,4-6]. In the UAE, it was clear that immediate action was required to transform health care delivery to a scalable digital system. [2,7]. Many hospitals and community clinics had to make a rapid transition from the previously limited scale of telemedicine to its widespread use as the primary mode of care delivery during the COVID-19 pandemic [8]. Telemedicine or telehealth, as defined by the World Health Organization, is the remote delivery of health care services and clinical information using digital technologies for the diagnosis, treatment, and prevention of disease and injuries and for the purposes of research, evaluation, and continuity of medical education [9]. During the COVID-19 pandemic, telemedicine consultations were primarily provided through (1) hospital outpatient departments (OPDs), which refer to moderately to highly integrated health care facilities providing secondary or tertiary health care services in a hospital setting, and (2) community clinics, which refer to ambulatory health care facilities that provide primary health care services. Because many of these health systems were implementing telemedicine technologies for the first time, it was unclear whether there were differences in the acceptance of this new technology for delivering health care services and whether patient satisfaction differed between the two health care systems. The acceptance of new technology was first described by Davis [10] in 1989 using the technology acceptance model (TAM). This model consists of two main constructs: (1) perceived usefulness and (2) perceived ease of use of the new technology. TAM can help us understand patients’ attitudes toward receiving clinical care through new online innovations such as telemedicine. The model also serves as a useful framework to understand intentions that influence the use of new technologies among the older generation. A more recent framework that resembled TAM was developed by Venkatesh and Davis [11] to unify technology

acceptance and use. The new model, referred to as the unified theory of acceptance and use of technology, incorporated perceived usefulness into a performance expectancy construct, perceived ease of use into an effort expectancy construct, and a social influence construct, which measures the effect of social factors on acceptance and use of new technology.

Patient satisfaction has been described by the US Centers for Medicare & Medicaid Services as the patient perspective of health care services, which can be used as an objective metric to compare hospitals and other health care organizations [12]. Patient satisfaction is becoming increasingly important in all aspects of health care [13]. It is a critical metric that is frequently used to assess the efficacy of health care services [14-17]. Thus, while patient satisfaction is a proxy, it is an effective way to measure the quality of health care services as published by the US Healthcare Effectiveness Data and Information Set report [18,19]. As with traditional health care systems, survey reports of patient satisfaction can help us understand patients’ attitudes toward telemedicine [20]. Moreover, by using reports of patient satisfaction with telemedicine, we can better understand patients’ experience with health care services, promote adherence to treatment [21,22], predict health care-related behaviors [23], and predict patterns of patients’ use of new health care technologies [24]. Community clinics are considered the primary source of care for patients seeking medical care in rural and urban communities [25,26], and previous reports have indicated that community clinics struggled with the rapid shift to telemedicine during the COVID-19 pandemic, unlike well-resourced hospitals that adapted swiftly to the new mode of care delivery [27]. Additionally, previously published studies revealed differences in satisfaction among patients receiving care in different settings, such as hospitals, community clinics, and physician offices [28-31]. Thus, we sought to explore the differences in patients’ perceptions of, and satisfaction with, telemedicine between hospital OPDs and community clinics, and to propose recommendations for future telemedicine delivery through these different systems using results from this survey study. We further aimed to explore patient- or visit-related factors that contribute to increased satisfaction with telemedicine and how these factors could be applied to quality assurance in the health care system.

## Methods

### Study Design

This was a survey-based cross-sectional study conducted in December 2020 on outpatients who used telemedicine services during the COVID-19 pandemic in Abu Dhabi. Data were collected using an online survey that was sent through an internal SMS system in a manner consistent with the American Association for Public Opinion Research reporting guidelines [32]. The study followed the STROBE (Strengthening the Reporting of Observational Studies in Epidemiology) reporting guidelines [33].

### Subject Selection and Inclusion and Exclusion Criteria

The sampling method used in this study was volunteer sampling. The calculated sample size to achieve 80% power was 426 participants, with a nonresponse rate of 20%. The online survey link was sent mainly via the internal SMS system of the Abu Dhabi Department of Health (DoH) and the Abu Dhabi Health Services Company (SEHA), which are two large health regulatory authorities in Abu Dhabi with a registry of patients who visited outpatient facilities (ie, hospital OPDs and community clinics) during the COVID-19 pandemic. The inclusion criteria for participants were as follows: (1) 18 years of age or older and (2) completed a telemedicine consultation in an outpatient setting during the COVID-19 pandemic from March to December 2020. We excluded patients who had never used telemedicine services during the COVID-19 pandemic.

### Survey Development, Piloting, and Data Collection

An online survey instrument was developed using questions from the validated Telemedicine Satisfaction Questionnaire and the Telemedicine Usability Questionnaire [34,35]. The survey consisted of demographic characteristics and eight questions, rated on a 5-point Likert scale, that were revised by a team of physicians who frequently consulted patients using telemedicine services during the COVID-19 pandemic; the survey was available in English and Arabic (Multimedia Appendix 1). Two main factors were examined in this survey: patient acceptance of telemedicine and patient satisfaction with telemedicine.

The first factor, patient acceptance of telemedicine, was examined through two main constructs:

1. Perceived usefulness of telemedicine, which was evaluated through three major dimensions: improvement of access to health care, saving time and costs, and addressing health needs.
2. Perceived ease of use of telemedicine, which was assessed through the following dimension: ease of expressing clinical concerns.

The second factor, patient satisfaction with telemedicine, was assessed through four dimensions:

1. Comfort during consultation.
2. Cultural compatibility.
3. Support for the transition to telemedicine.
4. Satisfaction with telemedicine.

A pilot study was conducted using a cohort of 30 patients to assess whether the questions were comprehensible, appropriate, well-defined, and understood in a consistent manner [36]. Each patient's information statement has also been evaluated for appropriateness and comprehension by the study investigators. The online survey instrument was built using the Microsoft Forms platform (version 2018; Microsoft Corporation). The survey was conducted over a 2-week period (ie, December 2-16, 2020); the initial invitation was sent in the first week followed by a reminder invitation in the second week to increase recruitment of subjects.

### Study Variables and Outcomes

Sociodemographic factors, including age, sex, education level, and marital status; modality of telemedicine (ie, video or audio consultation); experience with telemedicine; distance to health care facility; and type of health care system (ie, hospital OPDs or community clinics) were all self-reported by survey respondents. We compared the perceived usefulness and ease of use (ie, acceptance) of telemedicine services, as well as patient satisfaction with these services, between hospital OPDs and community clinics. We defined hospital OPDs as moderately to highly integrated health centers with high differentiation in the level of secondary and tertiary health services. We defined community clinics as ambulatory health practices that have a limited level of differentiation across services, providing mainly primary health care services. Differentiation was defined as the number of different services that the system is providing; integration has been measured by whether or not services are offered through this health system and whether or not physicians are aligned through a contractual mechanism [37,38]. These definitions were adapted from a widely recognized published taxonomy of health systems and networks by Bazzoli and colleagues [39]. Video consultation was defined as any remote clinical consultation taking place on a video platform using the camera in the patient's smartphone, tablet, or computer, where both the patient and physician can interact in a real-time manner [40]. Audio consultation was defined as any remote clinical consultation taking place through a phone call where interaction between the physician and patient is limited to an audio conversation only [8,11,41-43]. We defined perceived usefulness as how patients perceive the usefulness of the telemedicine system regarding improvement in the performance of, and access to, health care [44]. Perceived ease of use was defined as the degree to which a person believes that using a tool or a system would be easy and free of effort [44]. Patient satisfaction was defined as a subjective measurement that reflects the difference between patient expectation and the quality of telemedicine services they have received [45,46]. We measured these constructs using a questionnaire that reflects the core theme of perceived usefulness, ease of use, and satisfaction. Lastly, we defined middle-aged as being between 40 and 59 years of age [47].

### Statistical Analysis

Differences in the perception of telemedicine services and patient satisfaction between hospital OPDs and community clinics were investigated based on various outcome variables. While the first set of outcomes was related to the perceived

usefulness and ease of use of telemedicine services, the second set of outcomes assessed patients' satisfaction with telemedicine services and whether a difference existed between these two health care systems. Descriptive statistics characterizing the survey respondents were reported as frequencies and percentages for all variables. To compare responses to survey questions among health care system types, we performed chi-square statistical tests at a significance level of .05. We used ordered logistic regression analyses to investigate the association between health care systems, modality, and outcome variables (ie, perceived usefulness and ease of use of telemedicine and patient satisfaction), adjusting for confounding factors, such as sociodemographic characteristics. A forced-entry approach was adopted to consider the variance inflation factor (VIF) diagnostic to prevent unreliable estimates of coefficients and odds ratios (ORs) due to high correlations among predictor variables. Results showed that multicollinearity was not a concern in the final models (VIF=1.1). We also checked the distribution of the responses to questions, based on a 5-point Likert scale, and found limited observations, particularly toward the extreme negative and positive ends of the scale (ie, "strongly agree" and "strongly disagree"). Considering that the number of patients who selected "strongly agree" or "strongly disagree" was not large enough to permit a meaningful statistical analysis, we merged "strongly agree" and "agree" responses under a positive direction, and "strongly disagree" and "disagree" responses were merged under a negative direction. These two statements

were found to involve the same attitude continuum toward the question and were collapsed into "disagreement," "neutral," and "agreement," as has been done in similar previous studies [8,48,49]. Regression results were reported as ORs and 95% CIs, with  $P < .05$  demonstrating statistical significance. Statistical analyses were performed using Stata Statistical Software (version 16.1; StataCorp LLC).

### Ethical Approval

We obtained ethical approval for this study from the Institutional Review Board (IRB) of Khalifa University (protocol No H21-006-2020) and the Abu Dhabi COVID-19 Research IRB Committee of the Abu Dhabi Department of Health (IRB reference number DOH/CVDC/2020/1747). A waiver for informed consent was granted due to the deidentified nature of this study.

## Results

### Overview

A total of 515 patients completed the survey, of whom 33.4% (n=172) had a telemedicine consultation through community clinics, while the majority (n=343, 66.6%) had a telemedicine consultation through hospital OPDs. The sociodemographic descriptive characteristics of the two groups were summarized and compared (Table 1).



**Table 1.** Patient sociodemographic characteristics and descriptive statistics by health care system.

Variables	Participants, n (%) <sup>a</sup>			P value
	Hospital outpatient department (n=343)	Community clinic (n=172)	Total (N=515)	
Total	343 (66.6)	172 (33.4)	515 (100)	N/A <sup>b</sup>
<b>Sex</b>				
Male	138 (40.2)	91 (52.9)	229 (44.5)	.006 <sup>c</sup>
Female	205 (59.8)	81 (47.1)	286 (55.5)	
<b>Age range (years)</b>				
≤39	133 (38.8)	82 (47.7)	215 (41.7)	.07
40-59	169 (49.3)	78 (45.3)	247 (48.0)	
≥60	41 (12.0)	12 (7.0)	53 (10.3)	
<b>Education level</b>				
High school or equivalent	118 (34.4)	63 (36.6)	181 (35.1)	.88
Bachelor's degree or equivalent	172 (50.1)	83 (48.3)	255 (49.5)	
Master's degree, PhD, or equivalent	53 (15.5)	26 (15.1)	79 (15.3)	
<b>Marital status</b>				
Single	54 (15.7)	34 (19.8)	88 (17.1)	.47
Married	260 (75.8)	126 (73.3)	386 (75.0)	
Others (widowed or divorced)	29 (8.5)	12 (7.0)	41 (8.0)	
<b>Past experience with telemedicine</b>				
Never used	209 (60.9)	97 (56.4)	306 (59.4)	.32
Used	134 (39.1)	75 (43.6)	209 (40.6)	
<b>Modality</b>				
Audio consultation	311 (90.7)	163 (94.8)	474 (92.0)	.11
Video consultation	32 (9.3)	9 (5.2)	41 (8.0)	
<b>Employment status</b>				
Employed	224 (65.3)	129 (75.0)	353 (68.5)	.03
Unemployed	119 (34.7)	43 (25.0)	162 (31.5)	
<b>Distance to health center (minutes)</b>				
<30	265 (77.3)	140 (81.4)	405 (78.6)	.28
>30	78 (22.7)	32 (18.6)	110 (21.4)	

<sup>a</sup>Percentages in the "Total" row are based on the total number of participants (N=515), while percentages in all other rows are based on the total values in their respective column headings.

<sup>b</sup>N/A: not applicable; a statistical test was not performed on the total group.

<sup>c</sup>The P value for each group of variables is reported in the top row of each group.

## Sociodemographic Characteristics

Compared to patients who had telemedicine consultations via community clinics, patients who had telemedicine consultations via hospital OPDs were mainly female (205/343, 59.8% vs 81/172, 47.1%;  $P=.006$ ), were middle-aged (40-59 years: 169/343, 49.3% vs 78/172, 45.3%;  $P=.07$ ), had a college degree (bachelor's degree: 172/343, 50.1% vs 83/172, 48.3%;  $P=.88$ ), were married (260/343, 75.8% vs 126/172, 73.3%;  $P=.47$ ), had no previous experience with telemedicine (209/343, 60.9% vs 97/172, 56.4%;  $P=.32$ ), were unemployed (119/343, 34.7% vs 43/172, 25.0%;  $P=.03$ ), and lived far from the health center

(78/343, 22.7% vs 32/172, 18.6%;  $P=.28$ ). The majority of patients who used telemedicine services in community clinics were young, male, not married, and employed and had previous experience with telemedicine. Regarding the modality of telemedicine, audio consultation was used more frequently than video consultation in both settings, and the frequency of using video consultation was higher in hospital OPDs than in community clinics; however, this result was not statistically significant (32/343, 9.3% vs 9/172, 5.2%;  $P=.11$ ).

## Perceived Usefulness and Ease of Use of Telemedicine Services

Perceived usefulness and ease of use of telemedicine services was assessed using a multi-item approach. We assessed patients'

agreement with each statement using a 5-point Likert scale (Table 2). Perception of telemedicine usefulness and ease of use were equally high, with no statistically significant difference between the two settings.

**Table 2.** Comparison of survey responses regarding perceived usefulness and ease of use of telemedicine services provided by the health care system.

Statements and responses	Participants, n (%) <sup>a</sup>			P value
	Hospital outpatient department (n=343)	Community clinic (n=172)	Total (N=515)	
Total	343 (66.6)	172 (33.4)	515 (100)	N/A <sup>b</sup>
<b>Telemedicine improved access to clinical care</b>				
Disagree and strongly disagree	32 (9.3)	16 (9.3)	48 (9.3)	.55 <sup>c</sup>
Neutral	64 (18.7)	39 (22.7)	103 (20.0)	
Agree and strongly agree	247 (72.0)	117 (68.0)	364 (70.7)	
<b>Telemedicine saved time and travel costs</b>				
Disagree and strongly disagree	25 (7.3)	9 (5.2)	34 (6.6)	.29
Neutral	39 (11.4)	27 (15.7)	66 (12.8)	
Agree and strongly agree	279 (81.3)	136 (79.1)	415 (80.6)	
<b>Telemedicine can address patients' health care needs</b>				
Disagree and strongly disagree	39 (11.4)	18 (10.5)	57 (11.1)	.61
Neutral	67 (19.5)	40 (23.3)	107 (20.8)	
Agree and strongly agree	237 (69.1)	114 (66.3)	351 (68.2)	
<b>Medical concerns are easily expressed via telemedicine</b>				
Disagree and strongly disagree	34 (9.9)	12 (7.0)	46 (8.9)	.20
Neutral	51 (14.8)	35 (20.3)	86 (16.7)	
Agree and strongly agree	258 (75.2)	125 (72.7)	383 (74.4)	

<sup>a</sup>Percentages in the "Total" row are based on the total number of participants (N=515), while percentages in all other rows are based on the total values in their respective column headings.

<sup>b</sup>N/A: not applicable; a statistical test was not performed on the total group.

<sup>c</sup>The P value for each group of variables is reported in the top row of each group.

## Patient Satisfaction With Clinical Consultation

Similarly, patient satisfaction with telemedicine was assessed using a multi-item approach consisting of four

satisfaction-related statements, which were rated on a 5-point Likert scale (Table 3). Patient satisfaction with telemedicine services was equally high with no statistically significant difference between the two settings.

**Table 3.** Comparison of survey responses regarding overall satisfaction with telemedicine services and clinical consultations provided by the health care system.

Statements and responses	Participants, n (%) <sup>a</sup>			P value
	Hospital outpatient department (n=343)	Community clinic (n=172)	Total (N=515)	
Total	343 (66.6)	172 (33.4)	515 (100)	N/A <sup>b</sup>
<b>Felt comfortable consulting the physician using telemedicine services</b>				
Disagree and strongly disagree	30 (8.7)	21 (12.2)	51 (9.9)	.14 <sup>c</sup>
Neutral	59 (17.2)	38 (22.1)	97 (18.8)	
Agree and strongly agree	254 (74.1)	113 (65.7)	367 (71.3)	
<b>Telemedicine is a culturally appropriate way to receive health care services</b>				
Disagree and strongly disagree	33 (9.6)	19 (11.0)	52 (10.1)	.88
Neutral	81 (23.6)	40 (23.3)	121 (23.5)	
Agree and strongly agree	229 (66.8)	113 (65.7)	342 (66.4)	
<b>Support the transition to telemedicine services during and after the pandemic</b>				
Disagree and strongly disagree	33 (9.6)	15 (8.7)	48 (9.3)	.44
Neutral	67 (19.5)	42 (24.4)	109 (21.2)	
Agree and strongly agree	243 (70.8)	115 (66.9)	386 (69.5)	
<b>Satisfied with the quality of telemedicine services</b>				
Disagree and strongly disagree	31 (9.0)	18 (10.5)	49 (9.5)	.19
Neutral	59 (17.2)	40 (23.3)	99 (19.2)	
Agree and strongly agree	253 (73.8)	114 (66.3)	367 (71.3)	

<sup>a</sup>Percentages in the "Total" row are based on the total number of participants (N=515), while percentages in all other rows are based on the total values in their respective column headings.

<sup>b</sup>N/A: not applicable; a statistical test was not performed on the total group.

<sup>c</sup>The P value for each group of variables is reported in the top row of each group.

### Multivariate Analysis

In the multivariate model, the use of video consultation was significantly associated with increased perceived usefulness and ease of use of telemedicine. As compared to patients who had audio calls, patients who had video consultations were 3 times more likely to report that telemedicine improved access to health care services (OR 3.06, 95% CI 1.17-8.03;  $P=.02$ ), 5 times more likely to report that telemedicine reduced waiting times and travel costs (OR 4.94, 95% CI 1.15-21.19;  $P=.03$ ),

and 2.63 times more likely to report that telemedicine can address patients' medical needs (OR 2.63, 95% CI 1.13-6.11;  $P=.03$ ). There was no statistically significant association between the type of health care system (ie, hospital OPD vs community clinic) and patients' perceptions toward telemedicine (Table 4). Surprisingly, middle-aged patients were more likely to have higher perceived usefulness of telemedicine, indicating greater acceptance of the new technology-based model of care delivery.

**Table 4.** Adjusted multivariate analysis for perceived usefulness and ease of use of telemedicine.

Variables	Telemedicine improved access to clinical care		Telemedicine saved time and travel costs		Telemedicine addressed patients' health care needs		Telemedicine eased expression of medical concerns	
	OR <sup>a</sup> (95% CI)	P value	OR (95% CI)	P value	OR (95% CI)	P value	OR (95% CI)	P value
Health care system (hospital vs clinic)	1.20 (0.80-1.81)	.37	1.01 (0.63-1.64)	.94	1.05 (0.71-1.56)	.80	1.03 (0.67-1.56)	.91
Modality (video call vs audio call)	3.06 (1.17-8.03)	.02	4.94 (1.15-21.19)	.03	2.63 (1.13-6.11)	.03	2.19 (0.89-5.38)	.09
Sex (female vs male)	0.88 (0.58-1.33)	.54	1.31 (0.81-2.12)	.27	1.21 (0.80-1.82)	.36	1.06 (0.69-1.63)	.79
<b>Age range (years)</b>								
≤39 vs ≥60	1.48 (0.78-2.84)	.23	1.82 (0.87-3.82)	.11	1.32 (0.68-2.56)	.42	0.88 (0.42-1.83)	.73
40-59 vs ≥60	1.95 (1.02-3.71)	.04	3.03 (1.12-6.46)	.004	1.30 (0.68-2.50)	.43	1.14 (0.55-2.37)	.73
<b>Education level</b>								
College degree vs high school diploma	0.78 (0.50-1.21)	.27	0.79 (0.48-1.32)	.37	0.75 (0.49-1.16)	.20	0.76 (0.48-1.20)	.24
PhD degree vs high school diploma	0.99 (0.52-1.90)	.99	2.06 (0.88-4.84)	.10	1.07 (0.58-1.98)	.82	0.96 (0.50-1.85)	.90
<b>Marital status</b>								
Married vs single	0.87 (0.50-1.50)	.62	1.21 (0.66-2.20)	.54	1.07 (0.63-1.80)	.80	0.95 (0.55-1.66)	.87
Others vs single	1.08 (0.46-2.56)	.85	1.22 (0.46-3.28)	.69	1.09 (0.47-2.56)	.83	1.09 (0.43-2.75)	.85
Past experience with telemedicine (used vs never used)	0.93 (0.63-1.38)	.72	0.99 (0.63-1.59)	>.99	1.17 (0.79-1.71)	.44	0.77 (0.52-1.16)	.22
Employment status (employed vs unemployed)	1.22 (0.76-1.94)	.41	0.75 (0.43-1.33)	.33	0.93 (0.59-1.49)	.78	0.88 (0.53-1.45)	.67
Distance to health center (>30 min vs <30 min)	1.05 (0.65-1.68)	.85	0.89 (0.52-1.52)	.66	0.95 (0.60-1.49)	.82	0.85 (0.53-1.36)	.49

<sup>a</sup>OR: odds ratio.

Additionally, when compared to patients who had audio consultations, patients who had video consultations were 2.88 times more likely to support the transition to telemedicine services during and after the pandemic (OR 2.88, 95% CI 1.18-7.07;  $P=.02$ ) and 2.57 times more satisfied with telemedicine services (OR 2.57, 95% CI 1.04-6.33;  $P=.04$ ). Similarly, when compared to patients aged 60 years or older, middle-aged patients were 2 times more likely to be satisfied

with telemedicine services (OR 2.12, 95% CI 1.09-4.14;  $P=.03$ ). Additionally, when compared to employed patients, unemployed patients were more likely to be satisfied with telemedicine (OR 0.57, 95% CI 0.35-0.94;  $P=.03$ ). However, sex, education level, marital status, experience with telemedicine, and distance to health care center were not significantly associated with patient satisfaction with telemedicine services during the pandemic (Table 5).

**Table 5.** Adjusted multivariate analysis for patient satisfaction with clinical consultations.

Variables	Felt comfortable consulting the physician using telemedicine services		Telemedicine is a culturally appropriate way to receive health care services		Support the transition to telemedicine services during and after the pandemic		Satisfied with the quality of telemedicine services	
	OR <sup>a</sup> (95% CI)	P value	OR (95% CI)	P value	OR (95% CI)	P value	OR (95% CI)	P value
Health care system (hospital vs clinic)	1.44 (0.96-2.16)	.08	1.06 (0.72-1.57)	.76	1.11 (0.75-1.65)	.61	1.32 (0.88-1.98)	.18
Modality (video call vs audio call)	2.25 (0.96-5.26)	.06	1.58 (0.75-3.32)	.23	2.88 (1.18-7.07)	.02	2.57 (1.04-6.33)	.04
Sex (female vs male)	1.06 (0.70-1.60)	.79	1.11 (0.74-1.65)	.61	1.02 (0.68-1.53)	.93	1.06 (0.70-1.61)	.78
<b>Age range (years)</b>								
≤39 vs ≥60	1.57 (0.80-3.11)	.19	1.11 (0.58-2.13)	.74	1.63 (0.84-3.16)	.15	1.46 (0.75-2.85)	.27
40-59 vs ≥60	1.72 (0.88-3.36)	.11	1.20 (0.63-2.26)	.58	1.80 (0.94-3.46)	.08	2.12 (1.09-4.14)	.03
<b>Education level</b>								
College degree vs high school diploma	0.76 (0.48-1.20)	.24	0.71 (0.47-1.09)	.12	0.76 (0.49-1.19)	.23	0.67 (0.43-1.06)	.09
PhD degree vs high school diploma	0.89 (0.47-1.66)	.71	1.01 (0.54-1.88)	.98	0.96 (0.51-1.78)	.89	0.96 (0.50-1.83)	.90
<b>Marital status</b>								
Married vs single	0.90 (0.53-1.54)	.70	0.89 (0.53-1.50)	.66	1.03 (0.61-1.74)	.90	0.88 (0.51-1.52)	.65
Others vs single	1.32 (0.51-3.41)	.56	0.59 (0.26-1.30)	.19	1.25 (0.53-2.96)	.62	0.91 (0.37-2.25)	.84
Past experience with telemedicine (used vs never used)	1.01 (0.68-1.50)	.96	1.04 (0.72-1.52)	.83	1.08 (0.73-1.59)	.72	1.05 (0.70-1.56)	.83
Employment status (employed vs unemployed)	0.73 (0.45-1.18)	.20	0.92 (0.58-1.44)	.70	0.77 (0.48-1.24)	.29	0.57 (0.35-0.94)	.03
Distance to health center (>30 min vs <30 min)	0.79 (0.50-1.24)	.31	0.90 (0.58-1.40)	.63	1.06 (0.66-1.71)	.80	0.81 (0.51-1.30)	.38

<sup>a</sup>OR: odds ratio.

## Discussion

### Principal Findings

The front lines of medicine in many health care systems, including primary care clinics, were severely disrupted during the COVID-19 emergency. Despite the initial shock, many health systems were quick to adapt to the use of digital technologies; however, for some health systems, the transition has been smoother and faster than for others. During this crisis, telemedicine services have proven to be an integral part of the global public health response and showed capacity to act as a “safety net” for patients when properly reinforced [50-52]. In this paper, we have critically examined patients’ acceptance of telemedicine as a new technology for health care delivery and patient satisfaction with telemedicine services across two common health system types: hospitals and community clinics.

We have further explored patient characteristics and factors that predict satisfaction with telemedicine services. The Institute of Medicine recommends assessing the quality of health systems’ services either through patient satisfaction reports or through technical and professional assessment [52]; therefore, we used patient satisfaction survey results as a proxy to evaluate telemedicine quality across two types of health systems. Results from this study highlight three key findings: (1) there were no statistically significant differences in patient satisfaction between hospitals and community clinics regarding telemedicine services, (2) video consultation was significantly associated with increased patient satisfaction with telemedicine during the pandemic, and (3) being middle-aged was a significant predictor for patient satisfaction with telemedicine services, indicating higher acceptance of digital health among this age group.

Our first key finding suggests that perception of usefulness of telemedicine services, ease of use of these services, and satisfaction with these services were equally high among patients who had their telemedicine consultations in either hospitals or community clinics; this indicated similar quality of digital health services across these two types of health systems. Digital health innovations in community clinics have existed for some time, although the extent to which they are used vary greatly between countries. It is time to embrace these new technologies and increase the use of these innovations for patient management and follow-up, especially in community clinics, and to not fundamentally limit their use in integrated hospitals. Our results showed that 66.6% of all telemedicine consultations occurred mainly through hospitals, while only 33.3% occurred through community clinics. These findings highlight the need to increase the implementation and delivery of digital health innovations, particularly in health facilities, which are often considered the first point of contact for patients seeking medical care [53]. There is a need to build on the current status quo and accelerate the rollout of these digital technologies for routine use in primary health care settings, such as community clinics, to increase access to health care services.

Community clinics are a pivotal part of the public health system that could significantly improve access to health care services for the most vulnerable segments of the population if properly implemented within communities [26]. The pandemic has served as a catalyst to propel the use of telemedicine technologies into routine practice, and there is a significant amount of optimism surrounding this step. Results from recent studies in telemedicine showed that many patients with long-term chronic conditions prefer remote monitoring; thus, it is vital to opt for digital transformation of primary care services and follow-up care [53,54].

The second key finding from this paper is that video consultation, as compared to audio consultation, was significantly associated with improved perceived usefulness of telemedicine and higher levels of patient satisfaction. It comes as no surprise that patients favored video consultation over audio consultation, as it breaks the psychological barrier, eases guided remote physical examination of the patient, facilitates clinical decision-making, and eases expression of patients' concerns [55]. Moreover, the new generation of "digital native" patients are experienced in digital technologies and are comfortable communicating via virtual platforms, such as Skype, FaceTime, and Zoom [56]. Therefore, it is intuitive to introduce a telemedicine curriculum in medical schools and propose a model of education to effectively leverage telemedicine technologies and artificial intelligence in patient management [57,58]. Mainstreaming telemedicine and video applications in health systems could reduce health care disparities [59]. The surge in developing telemedicine applications with video call features is one of the most defining trends in this decade and will have a profound impact on socioeconomic and geopolitical realities, in particular in low- and middle-income countries (LMIC) [60,61]. With these telemedicine platforms, it is now possible to widen telemedicine use to remote geographical areas in LMIC and war zones. Policies advocating for the use of video consultation for certain patient categories should be implemented

at the grassroots level. Such policies can specify, for instance, regulations for acute and chronic patient management; they can also specify recommendations for new or follow-up patient care and whether it is recommended to have an in-person visit, video consultation, or audio consultation, depending on the initial evaluation using the Triage and Acuity Scale [62].

The third key finding from this paper is that middle-aged patients had a higher perceived usefulness and satisfaction with telemedicine when compared to patients in other age groups. In Abu Dhabi, at least one in five middle-aged patients showed acceptance of telemedicine use, possibly because telemedicine is convenient, safe, efficient, and cost-effective and can improve work-life balance [63,64]. However, we expected to see a satisfaction gap where Millennials—also known as Generation Y, born between 1981 and 1996—have higher levels of satisfaction, as they are labeled the "technology-savvy generation," relying heavily on technology and social media platforms for communication and addressing their life needs [65-67]. Our findings showed that patients belonging to Generation X—also known as the Baby Bust generation, born between 1960 and 1980—were the most satisfied with telemedicine. Thus, we propose that when studying patient experience with telemedicine, generational differences should be investigated further.

### Strengths and Limitations

This study has several strengths. To our knowledge, this is the first study to explore the differences in patient acceptance of, and satisfaction with, telemedicine between different health system types using patient survey results. Second, the study used a piloted and validated questionnaire that was derived from previously published studies in peer-reviewed journals. Third, the study measured the effect of telemedicine modalities (ie, video or audio consultation) on patient acceptance of, and satisfaction with, telemedicine, which is informative for decision-making policies.

Despite these strengths, the study has several limitations. First, this was a cross-sectional study capturing data entries at a single point in time with no comparison between in-person visits before and during the COVID-19 pandemic. Yet, while we felt that it was not possible to have an equal comparison since the number of in-person visits were very scarce during the pandemic due to the challenging situation, we plan to investigate this in future studies. Second, patient preference for telemedicine versus in-person visits was not captured, in addition to preference for video versus audio consultations. Third, the study did not investigate the factors influencing the age gap in telemedicine satisfaction and acceptance; however, we plan to address this as well in future studies. Moreover, this study did not measure the behavioral intention toward using video versus audio consultations. We aim to explore these factors in future studies. Lastly, our results could have been subjected to self-selection bias, as it is possible that patients who favored telemedicine or those who were more tech savvy were the ones who were motivated to participate in the study.

## Conclusions

During the COVID-19 pandemic, telemedicine played a positive role in improving health system and patient outcomes. While there are several studies in the literature that evaluated patient experience with telemedicine, there are no studies that evaluated the difference in patient satisfaction with telemedicine between different system types. Our study findings showed that patient satisfaction with telemedicine did not vary between different health care settings; however, further studies are needed to

objectively assess the differences in quality of telemedicine between these two systems. This study also demonstrated that video consultation was associated with higher patient satisfaction and improved teleconsultation experience. This finding may support and accelerate the rollout of video applications for all health care systems. Moving forward, it is vital to augment digital health innovations in community clinics in order to create a sustainable and effective health care system that is capable of coping with generational and technological trends.

## Acknowledgments

This work was endorsed by the Abu Dhabi Public Health Center and the DoH. We are grateful to Ms Amina Asghar, clinical research counselor from SEHA Corporate Academic Affairs, for assisting with the survey dissemination across Abu Dhabi SEHA hospitals. We are deeply thankful to Ms Mandy Chen, senior administrator in medical instructional design from Khalifa University College of Medicine and Health Science, for building the online surveys on Microsoft Forms.

## Authors' Contributions

NA and BA were responsible for conceptualization and design of the study and project administration. NA, BA, NB, MAA, RA, SAM, FAH, YAZ, and HA were responsible for the study investigation and data curation. MCES and NA were responsible for the formal analysis. NA, MCES, BA, EH, MA, HM, NB, MAA, RA, and AMA were responsible for preparation and writing of the original draft of the manuscript. NA, MCES, BA, EH, MA, HM, AMA, NB, MAA, RA, SAM, FAH, YAZ, SA, and HA reviewed and edited the manuscript. All authors have read and agreed to the published version of the manuscript.

## Conflicts of Interest

None declared.

## Multimedia Appendix 1

Patient survey.

[[DOCX File, 20 KB - medinform\\_v10i2e32373\\_app1.docx](#)]

## References

1. Blumenthal D, Fowler EJ, Abrams M, Collins SR. Covid-19 — Implications for the health care system. *N Engl J Med* 2020 Oct 08;383(15):1483-1488. [doi: [10.1056/nejmsb2021088](https://doi.org/10.1056/nejmsb2021088)]
2. Keesara S, Jonas A, Schulman K. Covid-19 and health care's digital revolution. *N Engl J Med* 2020 Jun 04;382(23):e82. [doi: [10.1056/NEJMp2005835](https://doi.org/10.1056/NEJMp2005835)] [Medline: [32240581](https://pubmed.ncbi.nlm.nih.gov/32240581/)]
3. Moonesar IA, Elsholkamy MM, Sayani H. The State of UAE Healthcare Service Delivery: Public Perceptions and Preliminary Insights. Dubai, UAE: UAE Public Policy Forum; 2018 Jan. URL: <https://mbrsgcdn.azureedge.net/cmsstorage/mbrsg/files/82/82dd2866-1197-4cc8-a3ce-0628733051d8.pdf> [accessed 2022-01-22]
4. Sanche S, Lin YT, Xu C, Romero-Severson E, Hengartner N, Ke R. High contagiousness and rapid spread of severe acute respiratory syndrome coronavirus 2. *Emerg Infect Dis* 2020 Jul;26(7):1470-1477 [FREE Full text] [doi: [10.3201/eid2607.200282](https://doi.org/10.3201/eid2607.200282)] [Medline: [32255761](https://pubmed.ncbi.nlm.nih.gov/32255761/)]
5. Monaghesh E, Hajizadeh A. The role of telehealth during COVID-19 outbreak: A systematic review based on current evidence. *BMC Public Health* 2020 Aug 01;20(1):1193 [FREE Full text] [doi: [10.1186/s12889-020-09301-4](https://doi.org/10.1186/s12889-020-09301-4)] [Medline: [32738884](https://pubmed.ncbi.nlm.nih.gov/32738884/)]
6. Croymans D, Hurst I, Han M. Telehealth: The right care, at the right time, via the right medium. *NEJM Catal Innov Care Deliv* 2020 Dec 30:1-12 [FREE Full text] [doi: [10.1056/CAT.20.0564](https://doi.org/10.1056/CAT.20.0564)]
7. Al-Zaman MS. Healthcare crisis in Bangladesh during the COVID-19 pandemic. *Am J Trop Med Hyg* 2020 Oct;103(4):1357-1359 [FREE Full text] [doi: [10.4269/ajtmh.20-0826](https://doi.org/10.4269/ajtmh.20-0826)] [Medline: [32828138](https://pubmed.ncbi.nlm.nih.gov/32828138/)]
8. Alhajri N, Simsekler MCE, Alfalasi B, Alhashmi M, AlGhatrif M, Balalaa N, et al. Physicians' attitudes toward telemedicine consultations during the COVID-19 pandemic: Cross-sectional study. *JMIR Med Inform* 2021 Jun 01;9(6):e29251 [FREE Full text] [doi: [10.2196/29251](https://doi.org/10.2196/29251)] [Medline: [34001497](https://pubmed.ncbi.nlm.nih.gov/34001497/)]
9. Ryu S. Telemedicine: Opportunities and developments in Member States: Report on the Second Global Survey on eHealth 2009 (Global Observatory for eHealth Series, Volume 2). *Healthc Inform Res* 2012;18(2):153. [doi: [10.4258/hir.2012.18.2.153](https://doi.org/10.4258/hir.2012.18.2.153)]
10. Davis FD. Perceived usefulness, perceived ease of use, and user acceptance of information technology. *MIS Q* 1989 Sep;13(3):319. [doi: [10.2307/249008](https://doi.org/10.2307/249008)]

11. Venkatesh V, Davis FD. A theoretical extension of the technology acceptance model: Four longitudinal field studies. *Manage Sci* 2000 Feb;46(2):186-204. [doi: [10.1287/mnsc.46.2.186.11926](https://doi.org/10.1287/mnsc.46.2.186.11926)]
12. HCAHPS: Patients' Perspectives of Care Survey. Baltimore, MD: US Centers for Medicare & Medicaid Services URL: <https://www.cms.gov/Medicare/Quality-Initiatives-Patient-Assessment-Instruments/HospitalQualityInits/HospitalHCAHPS> [accessed 2022-01-22]
13. Banka G, Edgington S, Kyulo N, Padilla T, Mosley V, Afsarmanesh N, et al. Improving patient satisfaction through physician education, feedback, and incentives. *J Hosp Med* 2015 Aug;10(8):497-502. [doi: [10.1002/jhm.2373](https://doi.org/10.1002/jhm.2373)] [Medline: [26014339](https://pubmed.ncbi.nlm.nih.gov/26014339/)]
14. Prakash B. Patient satisfaction. *J Cutan Aesthet Surg* 2010 Sep;3(3):151-155 [FREE Full text] [doi: [10.4103/0974-2077.74491](https://doi.org/10.4103/0974-2077.74491)] [Medline: [21430827](https://pubmed.ncbi.nlm.nih.gov/21430827/)]
15. Thomas AM. Patient satisfaction: Measuring the art of medicine. *JAMA* 1998 Dec 23;280(24):2127D. [doi: [10.1001/jama.280.24.2127](https://doi.org/10.1001/jama.280.24.2127)]
16. Al-Abri R, Al-Balushi A. Patient satisfaction survey as a tool towards quality improvement. *Oman Med J* 2014 Jan;29(1):3-7 [FREE Full text] [doi: [10.5001/omj.2014.02](https://doi.org/10.5001/omj.2014.02)] [Medline: [24501659](https://pubmed.ncbi.nlm.nih.gov/24501659/)]
17. Farley H, Enguidanos ER, Coletti CM, Honigman L, Mazzeo A, Pinson TB, et al. Patient satisfaction surveys and quality of care: An information paper. *Ann Emerg Med* 2014 Oct;64(4):351-357. [doi: [10.1016/j.annemergmed.2014.02.021](https://doi.org/10.1016/j.annemergmed.2014.02.021)] [Medline: [24656761](https://pubmed.ncbi.nlm.nih.gov/24656761/)]
18. Kruse CS, Krowski N, Rodriguez B, Tran L, Vela J, Brooks M. Telehealth and patient satisfaction: A systematic review and narrative analysis. *BMJ Open* 2017 Aug 03;7(8):e016242 [FREE Full text] [doi: [10.1136/bmjopen-2017-016242](https://doi.org/10.1136/bmjopen-2017-016242)] [Medline: [28775188](https://pubmed.ncbi.nlm.nih.gov/28775188/)]
19. D'Amore JD, Li C, McCrary L, Niloff JM, Sittig DF, McCoy AB, et al. Using clinical data standards to measure quality: A new approach. *Appl Clin Inform* 2018 Apr;9(2):422-431 [FREE Full text] [doi: [10.1055/s-0038-1656548](https://doi.org/10.1055/s-0038-1656548)] [Medline: [29898468](https://pubmed.ncbi.nlm.nih.gov/29898468/)]
20. Gustke SS, Balch DC, West VL, Rogers LO. Patient satisfaction with telemedicine. *Telemed J* 2000 May;6(1):5-13. [doi: [10.1089/107830200311806](https://doi.org/10.1089/107830200311806)]
21. Świątoniowska-Lonc N, Polański J, Tański W, Jankowska-Polańska B. Impact of satisfaction with physician-patient communication on self-care and adherence in patients with hypertension: Cross-sectional study. *BMC Health Serv Res* 2020 Nov 16;20(1):1046 [FREE Full text] [doi: [10.1186/s12913-020-05912-0](https://doi.org/10.1186/s12913-020-05912-0)] [Medline: [33198739](https://pubmed.ncbi.nlm.nih.gov/33198739/)]
22. Dubina MI, O'Neill JL, Feldman SR. Effect of patient satisfaction on outcomes of care. *Expert Rev Pharmacoecon Outcomes Res* 2009 Oct 09;9(5):393-395. [doi: [10.1586/erp.09.45](https://doi.org/10.1586/erp.09.45)] [Medline: [19817520](https://pubmed.ncbi.nlm.nih.gov/19817520/)]
23. Ware JE, Davies-Avery A, Stewart AL. The measurement and meaning of patient satisfaction. *Health Med Care Serv Rev* 1978;1(1):1, 3-1,15. [Medline: [10297474](https://pubmed.ncbi.nlm.nih.gov/10297474/)]
24. Manzoor F, Wei L, Hussain A, Asif M, Shah SIA. Patient satisfaction with health care services: An application of physician's behavior as a moderator. *Int J Environ Res Public Health* 2019 Sep 09;16(18):3318 [FREE Full text] [doi: [10.3390/ijerph16183318](https://doi.org/10.3390/ijerph16183318)] [Medline: [31505840](https://pubmed.ncbi.nlm.nih.gov/31505840/)]
25. Shin P, Sharac J, Jacobs F. Provision of telemedicine services by community health centers. *Online J Public Health Inform* 2014;6(2):e185 [FREE Full text] [doi: [10.5210/ojphi.v6i2.5421](https://doi.org/10.5210/ojphi.v6i2.5421)] [Medline: [25422721](https://pubmed.ncbi.nlm.nih.gov/25422721/)]
26. Rosenbaum S, Tolbert J, Sharac J, Shin P, Gunsalus R, Zur J. Community Health Centers: Growing Importance in a Changing Health Care System. San Francisco, CA: Henry J Kaiser Family Foundation; 2018 Mar 09. URL: <https://files.kff.org/attachment/Issue-Brief-Community-Health-Centers-Growing-Importance-in-a-Changing-Health-Care-System> [accessed 2022-01-22]
27. Kim JH, Desai E, Cole MB. How the rapid shift to telehealth leaves many community health centers behind during the COVID-19 pandemic. *Health Affairs Forefront*. 2020 Jun 02. URL: <https://www.healthaffairs.org/doi/10.1377/forefront.20200529.449762/full/> [accessed 2022-01-22]
28. Helliwell PS. Comparison of a community clinic with a hospital out-patient clinic in rheumatology. *Br J Rheumatol* 1996 Apr;35(4):385-388. [doi: [10.1093/rheumatology/35.4.385](https://doi.org/10.1093/rheumatology/35.4.385)] [Medline: [8624646](https://pubmed.ncbi.nlm.nih.gov/8624646/)]
29. Artiga S, Orgera K, Pham O. Disparities in Health and Health Care: Five Key Questions and Answers. San Francisco, CA: Henry J Kaiser Family Foundation; 2020 Mar. URL: <https://files.kff.org/attachment/Issue-Brief-Disparities-in-Health-and-Health-Care-Five-Key-Questions-and-Answers> [accessed 2022-01-22]
30. Hu R, Liao Y, Du Z, Hao Y, Liang H, Shi L. Types of health care facilities and the quality of primary care: A study of characteristics and experiences of Chinese patients in Guangdong Province, China. *BMC Health Serv Res* 2016 Aug 02;16(a):335 [FREE Full text] [doi: [10.1186/s12913-016-1604-2](https://doi.org/10.1186/s12913-016-1604-2)] [Medline: [27484465](https://pubmed.ncbi.nlm.nih.gov/27484465/)]
31. Forrest CB, Whelan EM. Primary care safety-net delivery sites in the United States: A comparison of community health centers, hospital outpatient departments, and physicians' offices. *JAMA* 2000 Oct 25;284(16):2077-2083. [doi: [10.1001/jama.284.16.2077](https://doi.org/10.1001/jama.284.16.2077)] [Medline: [11042756](https://pubmed.ncbi.nlm.nih.gov/11042756/)]
32. Pitt SC, Schwartz TA, Chu D. AAPOR reporting guidelines for survey studies. *JAMA Surg* 2021 Aug 01;156(8):785-786. [doi: [10.1001/jamasurg.2021.0543](https://doi.org/10.1001/jamasurg.2021.0543)] [Medline: [33825811](https://pubmed.ncbi.nlm.nih.gov/33825811/)]
33. von Elm E, Altman DG, Egger M, Pocock SJ, Gøtzsche PC, Vandenbroucke JP, STROBE Initiative. The Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) statement: Guidelines for reporting observational



- studies. *Ann Intern Med* 2007 Oct 16;147(8):573-577 [FREE Full text] [doi: [10.7326/0003-4819-147-8-200710160-00010](https://doi.org/10.7326/0003-4819-147-8-200710160-00010)] [Medline: [17938396](https://pubmed.ncbi.nlm.nih.gov/17938396/)]
34. Yip MP, Chang AM, Chan J, MacKenzie AE. Development of the Telemedicine Satisfaction Questionnaire to evaluate patient satisfaction with telemedicine: A preliminary study. *J Telemed Telecare* 2003;9(1):46-50. [doi: [10.1258/135763303321159693](https://doi.org/10.1258/135763303321159693)] [Medline: [12641893](https://pubmed.ncbi.nlm.nih.gov/12641893/)]
  35. Parmanto B, Lewis AN, Graham KM, Bertolet MH. Development of the Telehealth Usability Questionnaire (TUQ). *Int J Telerehabil* 2016;8(1):3-10 [FREE Full text] [doi: [10.5195/ijt.2016.6196](https://doi.org/10.5195/ijt.2016.6196)] [Medline: [27563386](https://pubmed.ncbi.nlm.nih.gov/27563386/)]
  36. Hassan ZA, Schattner P, Mazza D. Doing a pilot study: Why is it essential? *Malays Fam Physician* 2006;1(2-3):70-73 [FREE Full text] [Medline: [27570591](https://pubmed.ncbi.nlm.nih.gov/27570591/)]
  37. Luke RD. Taxonomy of health networks and systems: A reassessment. *Health Serv Res* 2006 Jun;41(3 Pt 1):618-628 [FREE Full text] [doi: [10.1111/j.1475-6773.2006.00524.x](https://doi.org/10.1111/j.1475-6773.2006.00524.x)] [Medline: [16704502](https://pubmed.ncbi.nlm.nih.gov/16704502/)]
  38. Piña IL, Cohen PD, Larson DB, Marion LN, Sills MR, Solberg LI, et al. A framework for describing health care delivery organizations and systems. *Am J Public Health* 2015 Apr;105(4):670-679. [doi: [10.2105/AJPH.2014.301926](https://doi.org/10.2105/AJPH.2014.301926)] [Medline: [24922130](https://pubmed.ncbi.nlm.nih.gov/24922130/)]
  39. Bazzoli GJ, Shortell SM, Dubbs N, Chan C, Kralovec P. A taxonomy of health networks and systems: Bringing order out of chaos. *Health Serv Res* 1999 Feb;33(6):1683-1717 [FREE Full text] [Medline: [10029504](https://pubmed.ncbi.nlm.nih.gov/10029504/)]
  40. Kitamura C, Zurawel-Balaura L, Wong RKS. How effective is video consultation in clinical oncology? A systematic review. *Curr Oncol* 2010 Jun;17(3):17-27 [FREE Full text] [doi: [10.3747/co.v17i3.513](https://doi.org/10.3747/co.v17i3.513)] [Medline: [20567623](https://pubmed.ncbi.nlm.nih.gov/20567623/)]
  41. Jaklevic MC. Telephone visits surge during the pandemic, but will they last? *JAMA* 2020 Oct 27;324(16):1593-1595. [doi: [10.1001/jama.2020.17201](https://doi.org/10.1001/jama.2020.17201)] [Medline: [33027503](https://pubmed.ncbi.nlm.nih.gov/33027503/)]
  42. Davis FD, Bagozzi RP, Warshaw PR. User acceptance of computer technology: A comparison of two theoretical models. *Manage Sci* 1989 Aug;35(8):982-1003. [doi: [10.1287/mnsc.35.8.982](https://doi.org/10.1287/mnsc.35.8.982)]
  43. Tamilmani K, Rana NP, Wamba SF, Dwivedi R. The extended unified theory of acceptance and use of technology (UTAUT2): A systematic literature review and theory evaluation. *Int J Inf Manage* 2021 Apr;57:102269. [doi: [10.1016/j.ijinfomgt.2020.102269](https://doi.org/10.1016/j.ijinfomgt.2020.102269)]
  44. Lin Z. The Overall Perception of Telemedicine and Intention to Use Telemedicine Services: A Comparison Between Frequent Travelers and Non-Frequent Travelers [master's thesis]. Ithaca, NY: Cornell University; 2017 Aug. URL: [https://ecommons.cornell.edu/bitstream/handle/1813/56813/Lin\\_cornell\\_00580\\_10117.pdf?sequence=1&isAllowed=y](https://ecommons.cornell.edu/bitstream/handle/1813/56813/Lin_cornell_00580_10117.pdf?sequence=1&isAllowed=y) [accessed 2022-01-22]
  45. Collins K, Nicolson P, Bowns I. Patient satisfaction in telemedicine. *Health Informatics J* 2016 Jul 25;6(2):81-85. [doi: [10.1177/146045820000600205](https://doi.org/10.1177/146045820000600205)]
  46. Linder-Pelz S. Toward a theory of patient satisfaction. *Soc Sci Med* 1982 Jan;16(5):577-582. [doi: [10.1016/0277-9536\(82\)90311-2](https://doi.org/10.1016/0277-9536(82)90311-2)]
  47. Lachman ME, Teshale S, Agrigoroaei S. Midlife as a pivotal period in the life course: Balancing growth and decline at the crossroads of youth and old age. *Int J Behav Dev* 2015 Jan 01;39(1):20-31 [FREE Full text] [doi: [10.1177/0165025414533223](https://doi.org/10.1177/0165025414533223)] [Medline: [25580043](https://pubmed.ncbi.nlm.nih.gov/25580043/)]
  48. Satin AM, Shenoy K, Sheha ED, Basques B, Schroeder GD, Vaccaro AR, et al. Spine patient satisfaction with telemedicine during the COVID-19 pandemic: A cross-sectional study. *Global Spine J* 2020 Oct 22:1-8 [FREE Full text] [doi: [10.1177/2192568220965521](https://doi.org/10.1177/2192568220965521)] [Medline: [33089712](https://pubmed.ncbi.nlm.nih.gov/33089712/)]
  49. Jacoby J, Matell MS. Three-point Likert scales are good enough. *J Mark Res* 1971 Nov;8(4):495. [doi: [10.2307/3150242](https://doi.org/10.2307/3150242)]
  50. Bhaskar S, Bradley S, Chattu VK, Adisesh A, Nurtazina A, Kyrkybayeva S, et al. Telemedicine across the globe-Position paper from the COVID-19 Pandemic Health System Resilience PROGRAM (REPROGRAM) International Consortium (Part 1). *Front Public Health* 2020;8:556720 [FREE Full text] [doi: [10.3389/fpubh.2020.556720](https://doi.org/10.3389/fpubh.2020.556720)] [Medline: [33178656](https://pubmed.ncbi.nlm.nih.gov/33178656/)]
  51. Maret D, Peters OA, Vaysse F, Vigarios E. Integration of telemedicine into the public health response to COVID-19 must include dentists. *Int Endod J* 2020 Jun;53(6):880-881 [FREE Full text] [doi: [10.1111/iej.13312](https://doi.org/10.1111/iej.13312)] [Medline: [32320488](https://pubmed.ncbi.nlm.nih.gov/32320488/)]
  52. Peabody JW, Taguiwalo MM, Robalino DA, Frenk J. Improving the quality of care in developing countries. In: Jamison DT, Breman JG, Measham AR, Alleyne G, Claeson M, Evans DB, et al, editors. *Disease Control Priorities in Developing Countries*. 2nd edition. Washington, DC and New York, NY: The World Bank and Oxford University Press; 2006:1293.
  53. Pagliari C. Digital health and primary care: Past, pandemic and prospects. *J Glob Health* 2021 Jul 02;11:01005 [FREE Full text] [doi: [10.7189/jogh.11.01005](https://doi.org/10.7189/jogh.11.01005)] [Medline: [34221352](https://pubmed.ncbi.nlm.nih.gov/34221352/)]
  54. Hanley J, Pinnock H, Paterson M, McKinstry B. Implementing telemonitoring in primary care: Learning from a large qualitative dataset gathered during a series of studies. *BMC Fam Pract* 2018 Jul 18;19(1):118 [FREE Full text] [doi: [10.1186/s12875-018-0814-6](https://doi.org/10.1186/s12875-018-0814-6)] [Medline: [30021535](https://pubmed.ncbi.nlm.nih.gov/30021535/)]
  55. Donaghy E, Atherton H, Hammersley V, McNeilly H, Bikker A, Robbins L, et al. Acceptability, benefits, and challenges of video consulting: A qualitative study in primary care. *Br J Gen Pract* 2019 Sep;69(686):e586-e594 [FREE Full text] [doi: [10.3399/bjgp19X704141](https://doi.org/10.3399/bjgp19X704141)] [Medline: [31160368](https://pubmed.ncbi.nlm.nih.gov/31160368/)]
  56. Haluza D, Naszay M, Stockinger A, Jungwirth D. Digital natives versus digital immigrants: Influence of online health information seeking on the doctor-patient relationship. *Health Commun* 2017 Nov;32(11):1342-1349. [doi: [10.1080/10410236.2016.1220044](https://doi.org/10.1080/10410236.2016.1220044)] [Medline: [27710132](https://pubmed.ncbi.nlm.nih.gov/27710132/)]

57. Pathipati AS, Azad TD, Jethwani K. Telemedical education: Training digital natives in telemedicine. *J Med Internet Res* 2016 Jul 12;18(7):e193 [FREE Full text] [doi: [10.2196/jmir.5534](https://doi.org/10.2196/jmir.5534)] [Medline: [27405323](https://pubmed.ncbi.nlm.nih.gov/27405323/)]
58. Aungst TD, Patel R. Integrating digital health into the curriculum-Considerations on the current landscape and future developments. *J Med Educ Curric Dev* 2020;7:1-7 [FREE Full text] [doi: [10.1177/2382120519901275](https://doi.org/10.1177/2382120519901275)] [Medline: [32010795](https://pubmed.ncbi.nlm.nih.gov/32010795/)]
59. Roghani A, Panahi S. Does telemedicine reduce health disparities? Longitudinal evidence during the COVID-19 pandemic in the US. medRxiv. Preprint posted online on March 2, 2021 [FREE Full text]
60. Mesko B. Health IT and digital health: The future of health technology is diverse. *J Clin Transl Res* 2018 Dec 17;3(Suppl 3):431-434 [FREE Full text] [Medline: [30873492](https://pubmed.ncbi.nlm.nih.gov/30873492/)]
61. Koonin LM, Hoots B, Tsang CA, Leroy Z, Farris K, Jolly T, et al. Trends in the use of telehealth during the emergence of the COVID-19 pandemic - United States, January-March 2020. *MMWR Morb Mortal Wkly Rep* 2020 Oct 30;69(43):1595-1599 [FREE Full text] [doi: [10.15585/mmwr.mm6943a3](https://doi.org/10.15585/mmwr.mm6943a3)] [Medline: [33119561](https://pubmed.ncbi.nlm.nih.gov/33119561/)]
62. Lin C, Tseng W, Wu J, Tay J, Cheng M, Ong H, et al. A double triage and telemedicine protocol to optimize infection control in an emergency department in Taiwan during the COVID-19 pandemic: Retrospective feasibility study. *J Med Internet Res* 2020 Jun 23;22(6):e20586 [FREE Full text] [doi: [10.2196/20586](https://doi.org/10.2196/20586)] [Medline: [32544072](https://pubmed.ncbi.nlm.nih.gov/32544072/)]
63. Kichloo A, Albosta M, Dettloff K, Wani F, El-Amir Z, Singh J, et al. Telemedicine, the current COVID-19 pandemic and the future: A narrative review and perspectives moving forward in the USA. *Fam Med Community Health* 2020 Aug;8(3):e000530 [FREE Full text] [doi: [10.1136/fmch-2020-000530](https://doi.org/10.1136/fmch-2020-000530)] [Medline: [32816942](https://pubmed.ncbi.nlm.nih.gov/32816942/)]
64. Shachar C, Engel J, Elwyn G. Implications for telehealth in a postpandemic future: Regulatory and privacy issues. *JAMA* 2020 Jun 16;323(23):2375-2376. [doi: [10.1001/jama.2020.7943](https://doi.org/10.1001/jama.2020.7943)] [Medline: [32421170](https://pubmed.ncbi.nlm.nih.gov/32421170/)]
65. Ng ESW, Schweitzer L, Lyons ST. New generation, great expectations: A field study of the millennial generation. *J Bus Psychol* 2010 Feb 16;25(2):281-292. [doi: [10.1007/s10869-010-9159-4](https://doi.org/10.1007/s10869-010-9159-4)]
66. Bolton RN, Parasuraman A, Hoefnagels A, Migchels N, Kabadayi S, Gruber T, et al. Understanding Generation Y and their use of social media: A review and research agenda. *J Serv Manage* 2013 Jun 14;24(3):245-267. [doi: [10.1108/09564231311326987](https://doi.org/10.1108/09564231311326987)]
67. Alkire (née Nasr) L, O'Connor GE, Myrden S, Köcher S. Patient experience in the digital age: An investigation into the effect of generational cohorts. *J Retailing Consum Serv* 2020 Nov;57:102221. [doi: [10.1016/j.jretconser.2020.102221](https://doi.org/10.1016/j.jretconser.2020.102221)]

## Abbreviations

**DoH:** Department of Health

**IRB:** Institutional Review Board

**LMIC:** low- and middle-income countries

**OPD:** outpatient department

**OR:** odds ratio

**SEHA:** Abu Dhabi Health Services Company

**STROBE:** Strengthening the Reporting of Observational Studies in Epidemiology

**TAM:** technology acceptance model

**UAE:** United Arab Emirates

**VIF:** variance inflation factor

*Edited by C Lovis; submitted 24.07.21; peer-reviewed by S Bidmon; comments to author 18.08.21; revised version received 26.08.21; accepted 21.12.21; published 15.02.22.*

*Please cite as:*

*Alhajri N, Simsekler MCE, Alfalasi B, Alhashmi M, Memon H, Housser E, Abdi AM, Balalaa N, Al Ali M, Almaashari R, Al Memari S, Al Hosani F, Al Zaabi Y, Almazrouei S, Alhashemi H*

*Exploring Quality Differences in Telemedicine Between Hospital Outpatient Departments and Community Clinics: Cross-sectional Study*

*JMIR Med Inform* 2022;10(2):e32373

URL: <https://medinform.jmir.org/2022/2/e32373>

doi:[10.2196/32373](https://doi.org/10.2196/32373)

PMID:[34978281](https://pubmed.ncbi.nlm.nih.gov/34978281/)

©Noora Alhajri, Mecit Can Emre Simsekler, Buthaina Alfalasi, Mohamed Alhashmi, Hamda Memon, Emma Housser, Abdulhamid Mustafa Abdi, Nahed Balalaa, Maryam Al Ali, Raghda Almaashari, Shammah Al Memari, Farida Al Hosani, Yousif Al Zaabi, Shereena Almazrouei, Hamed Alhashemi. Originally published in *JMIR Medical Informatics* (<https://medinform.jmir.org>), 15.02.2022. This is an open-access article distributed under the terms of the Creative Commons Attribution License

(<https://creativecommons.org/licenses/by/4.0/>), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Medical Informatics, is properly cited. The complete bibliographic information, a link to the original publication on <https://medinform.jmir.org/>, as well as this copyright and license information must be included.

---

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

---

<https://www.jmirpublications.com/>